Journal of
Sensor and Actuator Networks

MDPI

Review

Trends in Intelligent Communication Systems: Review of Standards, Major Research Projects, and Identification of Research Gaps

Konstantinos Koufos [1], Karim Haloui [1], Mehrdad Dianati [1,*], Matthew Higgins [1], Jaafar Elmirghani [2], Muhammad Imran [3] and Rahim Tafazolli [4]

1 Warwick Manufacturing Group (WMG), University of Warwick, Coventry CV4 7AL, UK; Konstantinos.Koufos@warwick.ac.uk (K.K.); Karim.el-Haloui@warwick.ac.uk (K.H.); M.Higgins@warwick.ac.uk (M.H.)
2 School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, UK; J.M.H.Elmirghani@leeds.ac.uk
3 James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK; Muhammad.Imran@glasgow.ac.uk
4 5G and 6G Innovation Centres, Institute for Communication Systems (ICS), University of Surrey, Guildford GU2 7XH, UK; r.tafazolli@surrey.ac.uk
* Correspondence: M.Dianati@warwick.ac.uk

Abstract: The increasing complexity of communication systems, following the advent of heterogeneous technologies, services and use cases with diverse technical requirements, provide a strong case for the use of artificial intelligence (AI) and data-driven machine learning (ML) techniques in studying, designing and operating emerging communication networks. At the same time, the access and ability to process large volumes of network data can unleash the full potential of a network orchestrated by AI/ML to optimise the usage of available resources while keeping both CapEx and OpEx low. Driven by these new opportunities, the ongoing standardisation activities indicate strong interest to reap the benefits of incorporating AI and ML techniques in communication networks. For instance, 3GPP has introduced the network data analytics function (NWDAF) at the 5G core network for the control and management of network slices, and for providing predictive analytics, or statistics, about past events to other network functions, leveraging AI/ML and big data analytics. Likewise, at the radio access network (RAN), the O-RAN Alliance has already defined an architecture to infuse intelligence into the RAN, where closed-loop control models are classified based on their operational timescale, i.e., real-time, near real-time, and non-real-time RAN intelligent control (RIC). Different from the existing related surveys, in this review article, we group the major research studies in the design of model-aided ML-based transceivers following the breakdown suggested by the O-RAN Alliance. At the core and the edge networks, we review the ongoing standardisation activities in intelligent networking and the existing works cognisant of the architecture recommended by 3GPP and ETSI. We also review the existing trends in ML algorithms running on low-power micro-controller units, known as TinyML. We conclude with a summary of recent and currently funded projects on intelligent communications and networking. This review reveals that the telecommunication industry and standardisation bodies have been mostly focused on non-real-time RIC, data analytics at the core and the edge, AI-based network slicing, and vendor inter-operability issues, whereas most recent academic research has focused on real-time RIC. In addition, intelligent radio resource management and aspects of intelligent control of the propagation channel using reflecting intelligent surfaces have captured the attention of ongoing research projects.

Keywords: intelligent networking; network slicing; network data analytics function (NWDAF); radio access network intelligent control (RIC)

Citation: Koufos, K.; Haloui, K.; Dianati, M.; Higgins, M.; Elmirghani, J.; Imran, M.; Tafazolli, R. Trends in Intelligent Communication Systems: Review of Standards, Major Research Projects, and Identification of Research Gaps. *J. Sens. Actuator Netw.* **2021**, *10*, 60. https://doi.org/10.3390/jsan10040060

Academic Editor: Lei Shu

Received: 17 August 2021
Accepted: 27 September 2021
Published: 12 October 2021

1. Introduction

The recent advent of heterogeneous services and use cases with diverse requirements have made modern wireless communication networks highly complex systems, which need to be carefully designed and operated to offer immersive experience to their customers, while keeping both capital expenditure (CapEx) and operational expenditure (OpEx) low. The increasing complexity of wireless ecosystems makes network planning, optimisation, and orchestration arduous tasks. Communication system engineers have made remarkable progress in modelling various digital and analogue transceivers and simple network functions, e.g., at the MAC and link layers in wired/wireless local networks. However, in the 5G/6G landscape, the traditional models for network optimisation often fall short, as they mostly enhance the performance of network processes or signal processing blocks one-by-one, with limited end-to-end (E2E) performance or cross-block considerations. Simulations can be useful in some cases, but their efficacy and applicability are often limited too.

New paradigms, ushered in by the recent developments in the field of artificial intelligence (AI), and, most specifically, data-driven machine learning (ML) techniques, offer new possibilities in operating complex systems, such as modern communication networks. Unlike traditional approaches, the aspiration is that a massive amount of data, which is now within our reach to collect and process within practical timeframes, can help alleviate the need for tedious mathematical modelling and simulations for operating complex systems. Deep neural networks (DNNs) and modern learning theories can enable computers to learn models and apply optimisations that are beyond human comprehension and intellectual capabilities. Not surprisingly, AI/ML is already accepted as an indispensable enabler for the closed-loop control and optimisation of several network processes in emerging wireless networks as recognised by several recent vision papers, such as [1–5].

Motivated by the aforementioned potentials, this article aims to provide the interested readers with a comprehensive analysis of the most recent progress in the use of data-driven and ML-based approaches in the study of modern communication systems and networks with a focus not only on the key, most recent publications, but also on the developments in the industry, major research projects and potential for new research presented by the gaps in the literature.

1.1. Related Survey Papers and Contributions of This Survey

We begin with an appraisal and acknowledgement of the related existing contributions in the literature related to AI/ML-empowered wireless communication networks. Our study and analysis of the field over several months revealed that the existing reviews usually do not investigate or contextualise research works through the lens of the ongoing standardisation effort, industrial take-ups, and collaborative research project frameworks. As a result, the taxonomy of various research contributions uses one of the following three identified methodologies: application-oriented [6–9], the traditional layering approach of the OSI protocol stack [10–17], or the learning method, i.e., supervised, unsupervised, DNNs, and reinforcement learning (RL) [18].

In this survey, we review the existing large-scale research activities on AI-enabled communications and networking from the perspective of standardisation activities to identify significant research gaps and inform and orientate future research directions to solve industrial needs. Even though this review may not be exhaustive, due to the fast development in this field, we believe that the suggested framework in this paper can be useful for future research initiatives, positioning articles, tutorial papers as well as future research efforts. It shall help researchers explain and position the scope of their work as part of a standardised intelligent networking architecture. The contributions of this survey are also summarised below.

- The O-RAN Alliance and the 3GPP standardisation body have already made significant progress in recommending how to perform data collection and intelligent control at the RAN and the core network, respectively. In terms of our approach, at the RAN,

we categorise ML-based technology components for transceiver design, according to their operational timescales as suggested by the O-RAN Alliance: real-time (less than 10 ms), near real-time (between 10 ms and 1000 ms), and non-real-time (larger than 1 s) intelligent control. For ML models operating at the edge and core networks, we summarize the main salient features of the network data analytics function (NWDAF) developed and standardised by 3GPP. It is worth noting that only a few research contributions are cognisant of the data collection and analytics architecture suggested by 3GPP.

- We highlight the use cases for AI-enabled communication networks recommended by ITU, the O-RAN Alliance and ongoing research projects to guide future research activities with high and significant business potentials.

1.2. Summary of the Paper

The remainder of this paper is organised as follows. In Section 2, we give a short review of intelligent functions at the RAN, and discuss purely data-driven and model-aided ML techniques as suggested by Renzo et al. [19]. In Section 3, we review the standardisation activities at the 5G core recommended by 3GPP and ETSI, and at the RAN by the O-RAN Alliance. We also discuss research and standardisation gaps identified by ITU FG-ML5G. Furthermore, we review the trends in ML algorithms running on low-power micro-controller units, often referred to as TinyML, which was mostly omitted by the previous survey papers. Section 4 provides a summary of recent and currently funded projects in Europe, the U.K., and the U.S. on intelligent communications and networking. Section 5 summarises the main research gaps identified during the combined review of research papers, projects, and standards, and concludes this survey.

2. Background on ML-Based Optimisation of Wireless Networks

The traditional approach for optimising the performance of wireless transceivers uses a divide-and-conquer concept. The transceiver is separated into signal processing blocks performing, e.g., channel equalisation, demodulation and error correction, which are optimised independently of each otherwith limited cross-block considerations [20]. For the optimisation of a signal processing block, well-established mathematical models are often used, derived from extensive research in information theory, signal processing, and statistics for the past many years [21]. Unfortunately, some of the models, despite being very accurate, can be mathematically intractable with a high complexity order for real-time operational use. For instance, MIMO detectors and pre-coders can be optimally designed based on a maximum a posteriori criterion, which, however, involves the solution of NP-hard problems. Hence, only sub-optimal techniques, such as zero forcing equalisation, can be pragmatically used [11]. Some tractable models may be overly simplistic and inadequate to accurately capture all the intricacies in the wireless propagation medium and the non-linearities and imperfections of hardware components [19]. Overall, the existing block-based transceiver structure associated with mathematical optimisation per block compromises the transceiver performance for simplicity and tractability, which AI/ML-based techniques aspire to address.

ML can be used to directly optimise the end-to-end performance by treating the entire system as a black box. This is usually referred to in the literature as the pure data-driven technique [19]. The most intuitive way to apply ML to wireless communications is to consider a transceiver as being an (unsupervised) auto-encoder constituted of an encoder and decoder. Then, both the transmitter and receiver can be implemented as feedforward neural networks (NNs) and can be jointly optimised as a single NN, while the communication channel is modelled by inserting extra hidden layers between them, e.g., a single layer adding white Gaussian noise is used in [22]. More complicated channel transfer functions including the tap-delay-line for wideband transmissions, frequency and phase offsets, and unknown time-of-arrival could be modelled by adding more layers [23]. The target in this approach is to train ML algorithms to learn the optimal system design,

i.e., the mapping of source symbols to transmitted waveforms along with optimal decoding at the receiver; see Figure 1.

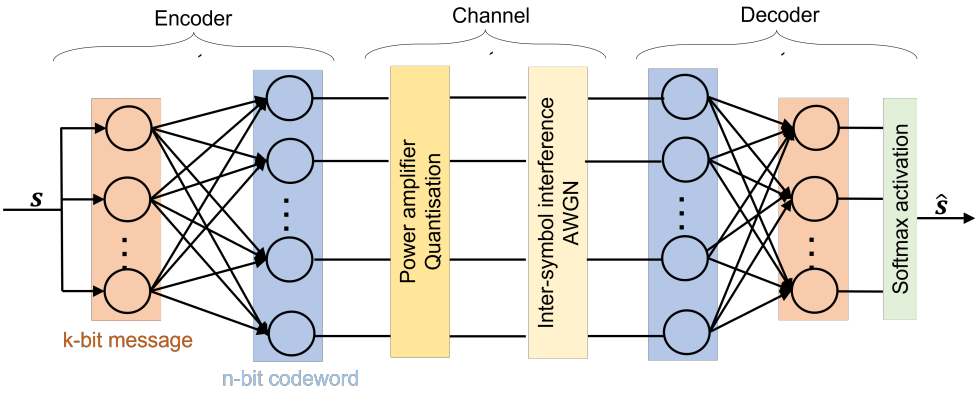

Figure 1. Block diagram of a channel auto-encoder with the transmitter and receiver implemented as fully-connected NNs and the channel incorporating linear effects, e.g., AWGN and inter-symbol interference, and non-linear distortions, e.g., power amplifier and quantisation. The output layer at the decoder applies a softmax activation function for soft symbol detection.

Implementing the transceiver as an end-to-end (E2E) auto-encoder is attractive, as it essentially alleviates the need for conventional signal processing without compromising performance; in [22,23], the block error rate in AWGN and multi-tap Rayleigh fading channel using auto-encoders is on par with the baseline schemes. Nevertheless, there are still significant unsolved challenges before adopting NN-based auto-encoders in practice, which are summarised below.

1. Training the auto-encoder over all possible source messages is required, which becomes quickly impractical for long code words. Additionally, it was observed that training at a low SNR does not necessarily generalise well at high SNRs [22], while training at multiple SNRs will prohibitively increase the size of required labelled datasets and time to train the NN.
2. The channel and all impairments between the transmitter and receiver must have a known deterministic functional form and be differentiable, which is seldom the case. For instance, the fading channel probabilistically varies over space and time. Furthermore, some impairments may not be differentiable, such as quantisation, or its mathematical representation may be inaccurate (e.g., the power amplifier response), or poorly understood (e.g., channel models for molecular and underwater communications). In this case, it is unclear how to backpropagate gradients from the receiver to the transmitter [24]. In addition, small discrepancies between the actual impairment and its model used for training may significantly degrade the performance during testing.
3. It is usually difficult to understand the relation between the topology of the NN, e.g., the number of layers, the activation and loss functions, and the performance of the transceiver. The explainability is a common problem hindering the adoption of AI/ML techniques in some application areas, though it may not be a significant obstacle in transceiver optimisation. Thus, efforts are needed not only to design new AI/ML models, but also to explain the working principle of the models. Fortunately, such research works are slowly emerging, for example, that of [25], which develops a parallel model to explain the behaviour of a recurrent neural network (RNN).

Due to the above limitations, many studies have already adopted a third approach for optimising the design of wireless systems, where ML and mathematical modelling work together and mutually benefit from each other. The main idea behind model-aided ML-based design is to keep the modular structure of the transceiver and use ML to optimise only some of the signal processing blocks, especially those involving complicated computations,

or for which some simplifications are assumed to make them mathematically tractable. See, for instance, in Figure 2, the model-assisted NN-based implementation of Viterbi decoding, where a small-sized NN is used to estimate the log-likelihood ratios at the receiver, given the channel output, while the rest of the detector follows the traditional Viterbi algorithm [26]. In the same direction, we list in Tables 1 and 2 some key studies, which are separated on the basis of operation timescales for RAN intelligent control (RIC), following the breakdown suggested by the O-RAN Alliance. Every research study comes along with a one-sentence summary of its main contribution.

Table 1. Example key studies utilising model-based ML techniques for improving the performance of wireless networks in real time (less than 10 ms).

Network Functions	Examples of Key Research Studies
Symbol detection	DNN-based receiver design for joint channel equalisation and symbol detection in OFDM systems [27].
	DNN-based MIMO detector combining deep unfolding with linear-MMSE for channel equalisation [28].
	LSTM-based learning of the log-likelihood ratios (LLRs) in Viterbi decoding [26].
Channel estimation	Estimating the channel state information (CSI) under combined time and frequency selective fading channels using DNNs [Yang2019] and convolutional neural networks (CNNs) [29].
	Adaptive channel equalisation using recurrent neural networks (RNNs) [30].
	Channel estimation using meta-learning [31].
	Reducing the CSI feedback in FDD massive MIMO systems using autoencoders in the feedback channel [32].
Channel prediction	DNN-based prediction of the downlink CSI based on the measured uplink CSI in FDD MIMO systems [33,34].
	RNN-based downlink channel prediction leveraging correlations in space and time to alleviate the issue of outdated CSI [35].
Channel coding	DNN-based decoding of short polar codes of rate ½ and block length N = 16 [36].
	DNN-based decoding of polar codes with length N = 128 leveraging the structure of belief propagation decoding algorithm [37].
	Integrating supervised learning for estimating the extrinsic LLRs into the max-log-map turbo decoder [38].
Link adaptation	Deep reinforcement learning (DRL)-based selection of the modulation and coding scheme (MCS) using the measured SNR as the environmental state and the experienced throughput as the reward [39].
	Supervised learning techniques for adaptive modulation and coding (AMC) including k-nearest-neighbours and support vector machines (SVMs) [40].
Reflecting intelligent surfaces (RIS)	DNN-based design and control of phase shifters [41].
	Using supervised learning to estimate the direct and the cascade (base station to RIS and RIS to the user) channels in RIS-based communication [42].
	Combining compressive sensing with deep learning (DL) to reduce the training overhead (due to the large number of reflecting elements) in the design of phase-shifts in RIS [43].
Spectrum sensing	Unsupervised (k-means, Gaussian mixture models) and supervised (k-nearest-neighbours and SVMs) learning methods for binary spectrum sensing [44].
	CNN-based multi-band cooperative binary spectrum sensing leveraging spatial and spectral correlations [45].
	Image-based automatic modulation identification based on the received constellation diagrams, signal distributions and spectrograms [46].

Table 2. Example key studies utilising model-based ML techniques for improving the performance of wireless networks in near real time (10 ms to 1000 ms), and non real time (longer than 1000 ms).

Operational Timescales	Network Functions	Examples of Key Research Studies
Near-real-time RIC	Resource allocation	DL for joint subcarrier allocation and power control in the downlink of multi-cell networks [47].
		Joint downlink power control and bandwidth allocation in multi-cell networks using NNs and Q-learning [48].
		DNNs for interference-limited power control that maximises the sum-rate under a power budget constraint at the base station [49].
		Predictive models on spectrum availability for distributed proactive dynamic channel allocation and carrier aggregation for maximising the throughput of LTE small cells operating in unlicensed spectrum bands [50].
		A DRL agent learns to select the best scheduling policy, including water-filling, round-robin, proportional-fair, or max-min, for each base station and RAN slice [51].
	Interference management	Centralised joint beam management and inter-cell interference coordination in dense mm-wave networks using DNNs [52].
Non-real-time RIC	E2E network slicing	Predicting the capacity of RAN slices and the congestion of network slices using NNs, and optimally combining them to sustain the quality-of-experience (QoE) [53].
		An AI/ML module predicts the states of network resources in runtime and autonomously stitches together RAN and core slices to satisfy all the user intents [54].
Network dimensioning and planning	SINR-based coverage evaluation using stochastic geometry	Deep transfer learning for selecting the downlink transmit power level that optimises the energy efficiency in cellular networks given the density of base stations [55].
		NN-based prediction of the coverage probability given the base station density, the propagation pathloss and the shadowing correlation model [56].
Network maintenance	Fault detection and compensation	Supervised learning method to detect performance degradation and the corresponding root cause of the fault [57].

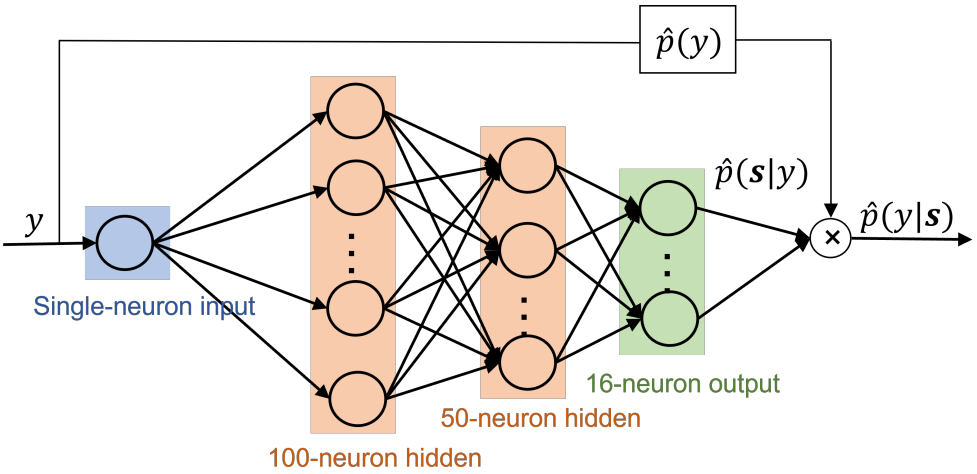

Figure 2. The Viterbi decoding algorithm has to estimate the log-likelihood ratio of the received symbol y for all possible transmitted sequences of symbols $\mathbf{s}$ to propagate along the Trellis diagram. Viterbinet uses a small-sized fully-connected NN with two hidden layers that is trained to estimate the likelihood for each possible sequence of transmitted symbols, given the channel output $p(\mathbf{s}|y)$. This is subsequently combined with the estimated marginal probability of the channel output $p(y)$ to calculate the desired log-likelihood ratios $p(y|\mathbf{s})$. The NN applies a softmax activation function at the output layer to calculate conditional probabilities. In the depicted NN, the output size is equal to 16 corresponding to BPSK modulated symbols transmitted over an intersymbol interference AWGN channel with memory length equal to 4 [26].

The summary of research works included in the tables reveals that the majority concentrate on real-time RIC, while, as we will shortly see in the next section, the industry and standardisation efforts have been mostly focused on the non-real-time RIC, data analytics at the core and the edge networks, AI-based E2E network slicing and vendor inter-operability issues. In our view, the joint creation of RAN and network slices, and the optimal proactive allocation of resources for E2E network slicing using AI/ML is a promising research topic worth pursuing.

3. AI/ML in the Standards and Industry

The first public deployments of 5G were a natural extension of the services offered by 4G by boosting connectivity through the enhanced mobile broadband (eMBB) use case, but the realisation of the full potential of 5G and beyond also stems from the support of new requirements offering lower latency services and the support of higher numbers of user equipment (UE) per unit surface area. Two new supported use cases, ultra-reliable and low-latency communications (uRLLC) and massive machine-type communications (mMTC), open new horizons to various industry verticals by providing new opportunities and prospects. Some representative examples of such verticals are healthcare, manufacturing, automotive, harbours, and retail.

The main technical enablers for providing customised services to verticals are mobile (or multi-access) edge computing (MEC) and network slicing [58]. British Telecom has calculated a 30% increase in revenue and a 40% decrease in OpEx by using only one physical infrastructure with network slicing instead of developing different physical networks to support various services. As we are gradually moving towards a standalone 5G network, a fully-fledged network slicing functionality brings a more granular control of the network resources. Capacity-based slices can be broken down into smaller pieces, offered to verticals and individual customers only for a few minutes or so. Network slicing also has an impact on the orchestration of the RAN, e.g., network and RAN slices may be jointly created for better E2E performance. The optimal operation of such a highly complex system across several geographical areas calls for an E2E automation with as little human intervention as possible.

Several standardisation bodies have already recognised that a fully automated network operation is closely linked to the effective use of big data analytics and AI/ML. AI-driven self-organised networking solutions by commercial vendors have also become increasingly popular [59]. Data collection engines could gather an enormous volume of raw data from several geographical areas and network locations (RAN/Edge/Core). The collected data can be subsequently analysed and mapped onto key performance indicators (KPIs), which are converted into appropriate actions at the RAN (e.g., radio resource allocation) and the core (e.g., intelligent network slicing) for satisfying the service level agreements (SLAs). Hidden patterns of the network behaviour can also be unveiled using AI/ML and provide the foundation for predicting interesting events, such as sharp load variations, connectivity outages, and faults. Additionally, industry verticals would require access to data analytics for assessing the acquired benefits against the induced cost.

Next, we review the existing works and future trends in the standardisation bodies and telecommunication industry related to the adoption of AI/ML and data analytics in 5G communication networks at the edge, the core and the RAN.

3.1. Data Analytics and AI/ML in 3GPP

Mobile network operators (MNOs) have traditionally relied on subscriber data collection (location, data rate, call drops, etc.) for network dimensioning and planning. The real-time monitoring of network failures has also been part of the operator's investment model, as it can improve anomaly detection and trigger (proactive) maintenance at reduced costs. With the advent of smartphones, we witnessed, for the first time, truly diversified services (data and voice), which created much more complicated data traffic patterns and made it imperative to adopt data analytics for network management and optimisation. Nevertheless, data analytics for enhancing the network performance have been so far diagnostic and mostly descriptive or, simply measuring various KPIs at several locations, which is suboptimal [60].

In 5G networks and beyond (B5G), not only are the offered services heterogeneous, but so are the end-devices. For instance, low-power Internet-of-Things (IoT) have very different communication requirements and hardware constraints than do autonomous vehicles. This makes E2E optimisation and the optimal allocation of network slices an arduous task with the added complexity that must be solved in real time and, thus, strengthening the need for big data analytics. Proactivity over reactivity would greatly benefit MNOs over various events, such as potential dynamic and sharp load changes (e.g., group mobility), upcoming outage events because of high interference in certain areas, network faults and maintenance. This amplifies the need to move from diagnostic and descriptive analytics to predictive analytics with an associated confidence level [61] and prescriptive data analytics with suggestions and antecedent countermeasures to maintain the required QoE levels [60]. 5G networks are designed with extended network function virtualisation (NFV) and MEC that allow the collection and processing of data across the network. 3GPP introduced in Release 15 the network data analytics function (NWDAF) as part of the 5G service-based architecture, a centralized entity for data collection and analytics at the core network [62]. The calculated analytics can either be predictive or provide statistics about past events.

3.1.1. Network Data Analytics Function (NWADF)

The NWDAF resides at the core network and utilises a bus architecture known as the service-based interface to communicate with other network functions (NFs). In the context of IMT-2020, a NF processes a certain network node functionality, e.g., session or mobility management, and has well-defined interfaces. In the cloud-native architecture, the virtualised NFs are broken down into smaller software units known as microservices, which are easier to manage and upgrade via open application programming interfaces (APIs) than large software components. Apart from NFs, the NWDAF can also provide analytics to application functions and the operation administration and maintenance (OAM) system. The interaction between the NWDAF and other NFs follows the consumer–producer model.

The NFs may subscribe to the NWDAF and request specific analytics (known as use cases or analytic events), which are given a unique identification number (ID) so that a NF can request a service from the NWDAF via its ID. The requests can be either periodic through a subscription to the NWDAF or ad hoc, i.e., one-off reporting request from the NWDAF.

NWDAF in Releases 15 and 16

In 3GPP Release 15, the consumers of the NWDAF are the policy control function (PCF) and the network slice selection function (NSSF); see Figure 3. The NWDAF receives load information from the NSSF and provides analytics that can be used to determine the optimal way to create or select the network slice(s) serving a UE or a group of UEs. The PCF uses NWDAF analytics to perform traffic steering to alleviate congestion at the core network and optimise the assignment of network resources [62]. In a nutshell, the NWDAF in Release 15 equips MNOs with data analytics to optimise and smartly manage several network slices, potentially running on the same physical infrastructure and instructing every few minutes or so the number of resources to allocate to each slice. This functionality is crucial to meet SLAs in a landscape with several offered services having diverged and/or conflicting latency, reliability, and data rate requirements. It is well understood that creating fine-grained slices and automating the allocation of network resources through AI/ML is imperative to ensure QoE per customer and service. Therefore, the NWDAF, where the AI/ML models reside, is a key component to achieve this goal. The alternative approach using large and relatively static slices will perform poorly as new services with diverse and stringent requirements appear in the market.

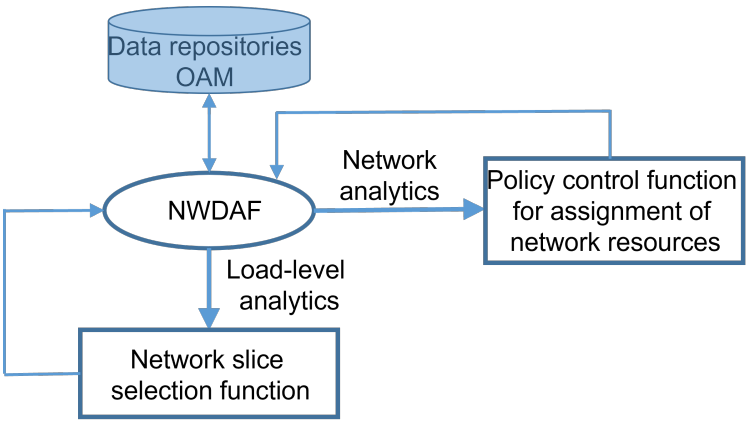

Figure 3. The NWDAF in 3GPP Release 15 for closed-loop optimisation of network slices.

In 3GPP Release 16, the NWDAF is allocated and dedicated to study and work items; it goes beyond closed-loop optimisation of network slices and interacts with many more NFs at the core network, e.g., access and mobility management (AMF) and session management (SMF) (see [63]). In principle, the NWDAF can exchange data with any NF in the core network, and correlate data received from multiple NFs to expose hidden patterns. NFs can also communicate with each other through the NWDAF. To give an example, the AMF supports (un)subscription to the notification of analytics information for SMF load information from NWDAF to improve SMF selection. Therefore, the NWDAF is both a producer and consumer in the service-based architecture. Overall, in Release 16, the NFs may provide local information to the NWDAF and request data analysis about various standardised events (using the unique ID of the event), including customised mobility management, mobility prediction, load balancing, QoS sustainability, service experience, and network performance; see [61] for more details on standardised analytic events.

While the communication interfaces (i.e., the APIs) between the NWDAF and other NFs are standardised by 3GPP, the implementation of specific AI/ML algorithms for big data analytics and closed-loop optimisation and automation are open to the needs of operators and vendors. Operators determine the location and the type of data and KPIs

to collect from the network, and specify the type of analytics the NWDAF provides to other NFs. In the B5G/6G landscape, it is envisioned that many vendors compete on the design of data collection techniques, ML model design, training, and inference, while the pipeline for communication and reporting back and forth to the NWDAF and other NFs is standardised. MNOs may also subcontract analytic services to several vendors, which increases market competition and innovation [64]. It should be noted that the NWDAF concerns analytics only in the core network; however, the separation between centralised and distributed units (CU/DU) in the RAN along with MEC have already set the stage for data analytics in the RAN and the edge, as we will shortly discuss below [65]. Currently, the NWDAF can access 3GPP RAN data only via the OAM entity. In this spirit, the study in [66] proposes an architecture that integrates analytics at the core network with application-level and RAN-centric analytics for E2E performance optimisation.

Another point worth mentioning is that similar NFs that need to communicate often with each other might be better provided by the same vendor, at least in the early deployments of 5G core networks. Similarly, an MNO may wish to subcontract the same supplier to build the core network at a specific geographical region. Although the service-based architecture congenitally supports a multi-vendor 5G core ecosystem (known as the best-of-breed model), the actual number of vendors providing the NFs might be limited in the beginning to ensure a viable operation.

Apart from the NWDAF, there is also the management data analytics function (MDAF), which is part of the operation and support system (OSS). It collects data from various repositories, the OAM system, and NFs, and provides analytics about network management and orchestration. The analytics calculated by the MDAF have a longer cycle than that collected by the NWDAF. Note that long-cycle analytics are not actionable immediately, but they can be used to identify new rules and expose hidden patterns that can eventually become a part of the NWDAF.

Not surprisingly, the available studies optimising wireless networks using AI/ML and that are compatible with the service-based architecture are limited. In [67], the authors consider a system composed of five cells, three categories of subscribers, five types of end-devices, and a realistic mobility model depending on the type of the device and the time of the day. They compare three ML models residing at the NWDAF for network load prediction and anomaly detection and demonstrate that RNNs and long-short-term-memory (LSTM) outperform linear regression methods. In [66], predictive data analytics are constructed by measuring the user mobility and the received signal strength in a vehicle-to-vehicle (V2V) communication scenario. The predictions are used to perform real-time radio resource management (RRM); it is shown that the session's throughput is increased in comparison to an uncoordinated allocation of resource blocks.

NWDAF in Release 17

According to the 3GPP specifications in Releases 15 and 16, multiple NWDAF instances, within the same public land mobile network (PLMN), act independently of one another. However, calculated analytics at nearby geographical areas and/or different analytics events may be related to each other, e.g., UE mobility and communication analytics are correlated with network data analytics. In addition, the sheer amount of generated data in future networks would rather impose a distributed architecture, where multiple NWDAF instances, placed at the edge and core of networks, cooperate with each other. Recall that the virtualised NFs are programmable and therefore, can be located almost anywhere in the network. Different NWDAFs may be assigned different tasks or share the computational load for the same analytic event. Furthermore, NWDAFs located at the edge can monitor the network, obtain useful analytics, and maintain the QoE, even if the connectivity to the central cloud is lost. Recall that communication to the cloud also increases latency; time-critical services, e.g., autonomous driving, are likely to require an NWDAF placed near the edge. For instance, an edge NWDAF might quickly perform

real-time analytics, while a centralised NWDAF is suitable for carrying out data processing, selection, and training of ML models.

In 3GPP Release 17, the functionality of NWDAF is also expanded towards specific use cases, i.e., 7 use cases and 21 key issues along with suggested solutions are described in [68]. As compared to Release 16, there are both data analytics and architectural enhancements. Firstly, the NWDAF is decomposed into two logical entities performing ML training and ML inference, respectively. Secondly, a hierarchical architecture is recommended, where a centralized NWDAF coordinates multiple localised NWDAFs [68] (p. 251). Unfortunately, the communication between NWDAFs raises privacy and security concerns. For example, consider the scenario where a NWDAF collects the trajectories of UEs and forwards them to the central NWDAF for training ML models. In this scenario, attack surfaces are created both at the NWDAFs and in their communication channel. To mitigate the risk of data breaches and cyber security attacks, it might be beneficial to perform local training at the point of data collection and transmit only the trained model parameters to the central NWDAF, which can aggregate several received models and send the outcome back to the learners. 3GPP Release 17 has suggested involving federated learning instead of sharing raw training data between NWDAF instances; see [68] (p. 133) for the general procedure of trained model registration, discovery, update, and consumption. See also an example illustration in Figure 4. Federated learning is shortly discussed next.

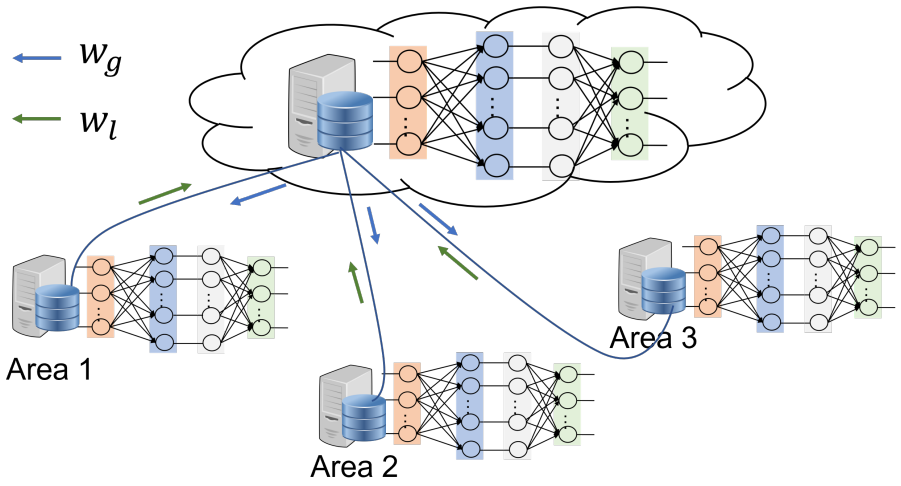

Figure 4. Example illustration for distributed implementation of NWDAF in 3GPP Release 17, using federated learning [68]. Three distributed NWDAF instances at the edge, occupying different geographical areas, perform local learning and share their NN weights $w_l, l \in \{1,2,3\}$ with the centralised NWDAF at the cloud, which averages the received models and sends back to the local learners the global model w_g for re-training.

Federated Learning

Federated learning is a distributed learning algorithm, where several learners are locally trained using their own datasets and communicate only their model parameters to a server. The server aggregates the received models to form the global model that is sent back to the learners for retraining. This process iteratively continues until a convergence constraint is satisfied. Federated learning was invented by Google for image classification and language modelling tasks and showed remarkable performance in next word prediction [69]. One should not confuse federated learning with other distributed learning methods, e.g., split learning [70], where clients with low computational capabilities, such as IoT, train only the first few layers of a DNN and send the extracted features to a gateway or to the cloud, which completes the computationally demanding part of the training on behalf of the clients.

Federated learning can become a bandwidth-efficient and privacy-preserving distributed learning technique for communication networks. The datasets collected by the

learners can vary in size and do not need to be independent of each other, which is often the case in communication networks with the learners being either distributed NFs at the core or UEs at the RAN. In [71], federated learning is used for predicting popular video content for edge caching and is performed closely to the centralised baseline scheme. The 5G core network has an inherent built-in support for distributed learning. Specifically, the various NFs can be viewed as learners that collect data and locally train ML models about specific analytic events, e.g., UE mobility, communication, and session management, while the NWDAF can be viewed as the server that supports global analytics, e.g., learning to optimise the allocation of network slices. Furthermore, in 3GPP Release 17, multiple NWDAF clients can form a federation and train ML models in a distributed and hierarchical manner [68]; see Figure 4.

Federated learning significantly reduces the communication signalling overhead as compared to the case where raw data are exchanged between the server and the learners. However, the number of model parameters, i.e., the number of NN weights, can also become large in the case of DNNs, not to mention the possibly many iterations required until convergence in federated learning. To further improve learning efficiency, federated distillation shares only the model output, i.e., the last layer of the DNN instead of the complete model, yielding, for example, 26 times less communication overhead at the cost of negligible performance loss in [72]. Other methods to reduce the signalling overhead involve model compression or model updates at longer intervals.

The study in [73] considers federated learning at the RAN, where the clients are UEs and the server is a gNB. To reduce the number of errors in the received model parameters, the gNB optimises the allocation of RBs to a selected subset of UEs that experiences favourable channel conditions. For federated learning at the RAN, analogue transmissions of the model parameters from the learners to the server can also be helpful. In the analogue domain, every learner can participate in the model updates, utilising the full available bandwidth to communicate with the server, and thus, the model convergence time decreases. Recall that, in federated learning, the server is not interested in the individual parameters calculated from each learner, but only in their summation, which is equal to the superposition of the received analogue signals at the server [74]. An extensive review of federated learning for MEC can be found in [75], while the study in [76] presents open research topics and future directions for federated learning in 6G communication networks.

3.1.2. European Telecommunications Standards Institute (ETSI)

Within 3GPP, the ETSI has already established several committees working on AI/ML-based operation for 5G/B5G wireless networks to enhance their automation and autonomy levels [77]. We discuss below the activities for two of these committees; the former is working towards the experiential networking intelligence (ENI) system, and the latter is producing the zero-touch network and service management (ZSM) architecture.

ETSI ENI

In traditional communication networks, network operation, optimisation and management are costly and time-consuming tasks exacerbated by vendor inter-operability issues and the recent advent of personalised services. The ENI initiative, founded in 2017, designs an AI-based cognitive network management system that provides feedback, recommendations, or commands to another system, e.g., to an MNO, an orchestrator, a user, or an application, aiming at improving the other system's performance and increasing its autonomy level in a cost effective manner [78]. The so-called assisted system may or may not have AI/ML functional modules, and most importantly, it does not need to change its functionality to receive help from the ENI system, as illustrated in Figure 5. To accelerate its adoption, the ENI system comes with an API broker that acts as a gateway and translates between the APIs of the two systems (assisted and ENI). The ENI system comprises six functional blocks, including situation-awareness and context-awareness functions (which enable it to dynamically adapt to changes in the environment), the assisted system's man-

agement objectives, business goals, and user needs. Essentially, the ENI and the assisted systems together create a closed loop so that the ENI system always receives updated data from the assisted system, and in turn, provides AI/ML-driven recommendations for network management, orchestration, etc.

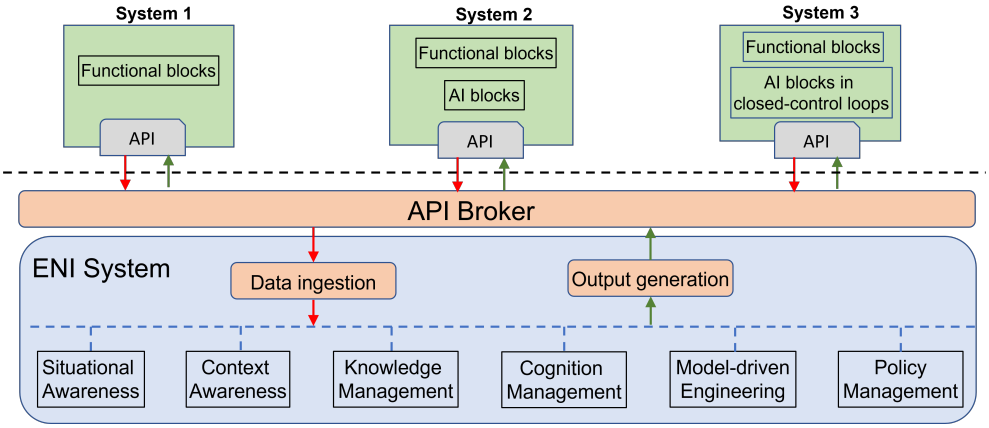

Figure 5. High-level functional architecture of the ENI system with an API broker and three assisted systems with different levels of AI adoption, ranging from no AI capability to AI functional blocks with closed-loop controls [79].

The ENI committee has identified several use cases where AI seems to have a high potential in improving network operation and management, and grouped them under five main categories, specifically, infrastructure management, network assurance, network operations, service orchestration and management, and network security [80]. Example representative use cases for each category are energy optimisation, network fault prediction, radio coverage and capacity optimisation, intelligent caching based on content popularity predictions, and policy-based network slicing for IoT security, respectively. For instance, the ENI system can interact through standardised interfaces (reference points) with an MNO, collect data and learn about the MNO's operation and performance, and finally, suggest methods to optimise the MNO orchestration of network slices. The recommendations can be dynamically updated as the ENI system collects more data, the environment changes or the goals of the MNO vary in time. Subsequently, the selected use cases by the ENI committee are associated with the requirements that the ENI system must fulfil [81]. To give an example, a functional requirement shall specify how to perform data collection, analysis, and learning, or how to inter-work with other systems, including the assisted system, while a network requirement shall determine how to allocate network resources to the virtual NFs of the assisted system.

In a nutshell, the ENI system is a versatile, AI/ML-based system with generic architecture that can support various decision-making processes and systems to achieve their goals, as specified in the ENI use cases. There are ongoing ENI system demonstrations (proofs-of-concept) [82], while the latest positioning white paper produced by the ENI committee is available in [83].

ETSI ZSM

The ZSM committee is another cross-industry initiative, launched by ETSI in 2017, whose aim is to design a reference architecture for future communication networks, that are solely managed by AI/ML without any human intervention [84]. The overarching goal of the ZSM committee is to design a reference architecture that supports self-monitoring, self-optimisation, and self-healing of future networks, with higher levels of autonomy attained as more data is collected [85]. The main drivers for this are the high complexity of future networks, the service-based architecture along with the adopted principles of SDN/NFV, the multi-vendor ecosystems, and the unprecedented business opportunities created by network slicing. The ZSM committee recognises that data collection and closed-

loop operations throughout the network are the lifeblood of network automation, see Figure 6.

Network automation and agility would allow MNOs to provide new services and business opportunities, which ultimately is beneficial to their customers, rather than managing the operation of their networks themselves.

To accelerate the adoption of E2E automation in multi-vendor environments, a single, harmonised, and interoperable service and network management framework is needed, where different solutions provided by vendors, standardisation bodies, and open-source projects can consistently fit together and inter-operate [86]. To achieve this target, the ZSM committee does not intend to generate a new standard but rather integrate existing standards and solutions to a consistent architectural framework. The system architecture suggested by the ZSM committee is modular, scalable, and service based with open, intent-based interfaces to facilitate the adoption of various existing solutions and standards [87].

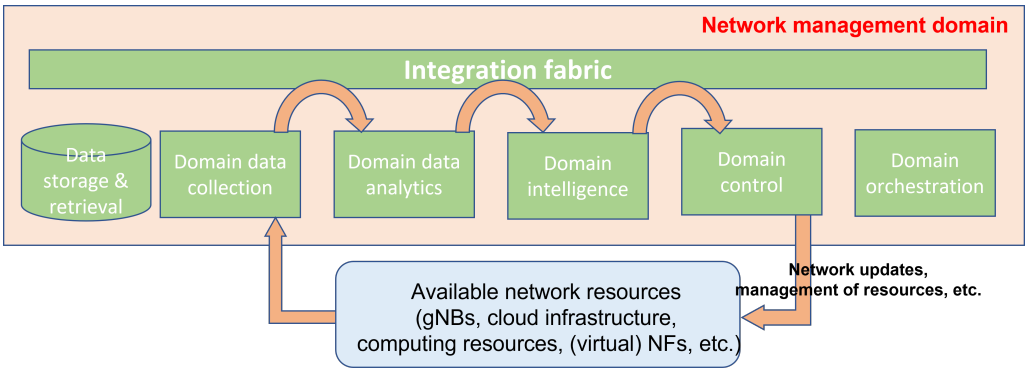

Figure 6. High-level reference architecture for a network management domain with closed-loop control and five domain functions as suggested by the ZSM committee. Domain-data analytics, e.g., for fault identification, are stored, and analytics about relevant past events are retrieved and enhance the performance of AI/ML tools residing under domain intelligence. Finally, the integration fabric is used as an interface for the exchange of information between different network management domains. Note that there are E2E services that consist of several domains, e.g., a service for E2E network slicing comprises core, transport, and RAN slice subnets spanning different management domains [87].

There are several other committees under the ETSI auspices with a focus on AI/ML-based networking solutions. For example, the Securing AI (SAI) group, formed in 2019, studies threats arising from the deployment of AI systems and aspires to provide mitigation schemes against cyber security attacks targeting AI systems. Another example is the Core Network and Inter-operability (INT) committee, which considers AI-aided actions to configure network parameters, e.g., routing policies. All these committees are likely to collaborate, develop synergies with standard developing organisations (SDOs) outside ETSI, and liaise with open-source projects, such as the Linux Acumos and the Open Network Automation Platform (ONAP) to avoid duplications and jointly realise the vision of digital industry transformation.

3.2. International Telecommunication Union (ITU)

The ITU-T study group (SG) 13 established the focus group (FG) on ML for future networks including 5G (FG-ML5G) in 2017, which was active between January 2018 and July 2020, and produced 10 technical specifications for the integration of ML into future networks [88]. The FG identified standardisation and research gaps but also specified data formats to be collected and processed for a variety of use cases. The use cases were grouped under five main categories, and for each of them, the requirements related to data collection, storage, and processing were determined. The identified use cases are relevant topics for further research work and are listed in Table 3.

Table 3. Identified use cases for ML in future networks by FG-ML5G [89].

Categories	Use Cases
Network slice and other network service-related use cases.	• Cognitive heterogeneous networks and ML-based self-optimised networks. • Radio resource management (RRM) for network slicing. • E2E network operation automation. • Service design • Network resource adaptation • Logical network design and deployment • Fault detection and recovery • Application-specific network slicing through in-network ML. • Smart traffic mirror—an ML-assisted network service. • ML-based E2E network slicing for 5G. • ML-based utility maximisation of sliced backhauls. • Energy efficient trusted multi-tenancy in IMT-2020 cross-haul. • Network slice SLA assurance based on ML. • Service management for smart cities. • Automated testing of services.
User-plane related use cases.	• Traffic classification. • Long-term traffic forecasting. • Emergency services based on ML.
Application related use cases.	• Access network assisted transmission control protocol window optimisation. • Retention and storage intelligence function. • Data-driven architecture for ML at the edge.
Signalling or management related use cases.	• ML mobility pattern prediction. • Load balance and cell splitting/merging. • ML-based QoE optimisation. • ML-based network management for Industry 4.0. • ML-based correlations between transport KPIs and radio KPIs. • ML-based E2E network management. • ML-aided channel modelling and channel prediction. • ML-based link adaptation optimisation.
Security-related use cases	• Combatting use of counterfeit ICT devices—ML-assisted network service. • ML-based identification of illegal exchanges using SIM boxes.

Another core contribution of the FG-ML5G is the recommendation for a high-level and interoperable architectural framework, which incorporates ML into future wireless networks [90]. In this architecture, a telecommunication network, e.g., 5G or IEEE 802.11, can become an underlay that provides data to the ML pipeline (or ML overlay). The latter collects the data through a standardised interface (the SRC) and passes them through a well-defined configuration, including data pre-processing and NNs associated with the subject use case. Finally, the ML overlay distributes the ML output to the corresponding SINK (another standardised interface), through which the output is applied to the underlay network. Apart from the SRC and SINK interfaces that interact with the NFs of the underlay network, the ML functionalities in the overlay are technology agnostic. Therefore, AI/ML algorithms and telecommunication technologies can evolve in parallel. The study in [91] utilises this architecture for ML-based association and handover in a WLAN.

Another critical architectural component suggested by the FG-ML5G is the ML sandbox, where ML models are trained and tested before their live deployment in the underlay network [90]. In the sandbox, MNOs can compare the performance of various ML models and mitigate the risks associated with the adoption of ML techniques. Training ML models in the sandbox may use either real-network or simulation data, leveraging digital twins of

the underlay network [92]. Therefore, the sandbox enables the ML pipeline to dynamically adapt to changes in the underlay network or the environment. The selected ML models are finally stored in a repository known as the ML marketplace, which is managed by an orchestrator and is responsible for selecting the best ML model according to the needs of the subject use case. For a high-level overview of the interactions among the architectural components recommended by the FG-ML5G, see Figure 7.

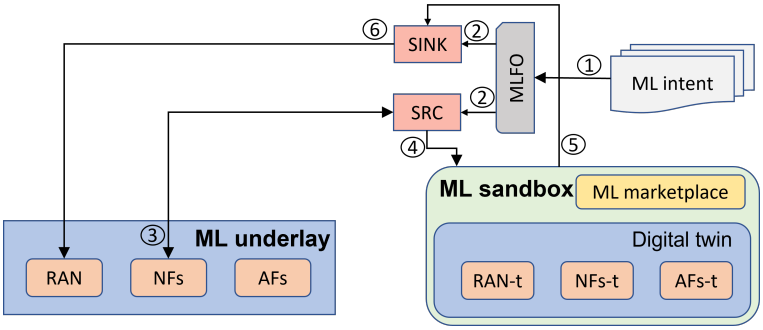

Figure 7. The ML intent is provided to the Machine Learning Function Orchestrator (MLFO), which specifies the associated SRC and SINK interfaces. In this illustration, the SRC is located at the core network and collects measurements, e.g., for access and mobility management (AMF) for a UE, which are subsequently provided to the ML sandbox. The sandbox retrieves suitable ML models from the marketplace, e.g., for mobility prediction, and assess their performance using a digital twin of the underlay network. AFs-t, NFs-t and RAN-t stand for the digital twin application functions, network functions and RAN implementation. Finally, the SINK which, in this illustration, is located at the RAN may for instance request the gNB to allocate more resource blocks to the UE as the latter is about to travel across an area with poor line-of-sight wireless propagation conditions.

During its lifetime, the FG-ML5G had a strong interaction with MNOs, with the ETSI ENI committee for creating a framework for the evaluation of intelligence levels [93], with the ETSI ZSM committee for designing the ML pipeline architecture, etc. Additionally, the FG encouraged the use of open-source frameworks and AI platforms, such as the Linux foundation for AI Acumos [94]. To engage the research community, the FG organised a global AI challenge between March and December 2020, where research teams competed to solve real telecommunication industry problems using AI/ML, e.g., network topology optimisation, and PHY channel estimation [95]. The ITU-T SG13 focus group on autonomous networks (FG-AN) was established in December 2020 and is the natural continuation of FG-ML5G [96].

3.3. Open RAN

We have seen in the previous section that the disaggregation between hardware and software along with programmability using the principles of SDN/NFV are crucial for adopting a service-based architecture (SBA) in the 5G core network, and subsequently integrating big data analytics and intelligent control. Open RAN aims at bringing intelligence also into the RAN and is mainly driven by the combined efforts of two organizations: the O-RAN Alliance, founded in 2018, focuses on the development of open and standardised interfaces in the RAN [97], while a project group inside the Telecom Infra Group (TIP) focuses on the design of a software-based RAN using commercial off-the-shelf (COTS) hardware and disaggregated software [98].

Using the principles of SDN/NFV, commodity hardware is used in the Open RAN (the O-RAN Alliance uses the term white box hardware), and the various functions are implemented in software, possibly by different vendors, due to openness. For instance, a remote radio unit (RRU) by vendor A can communicate via an open interface to a virtual PHY or MAC layer function developed by vendor B running on an Intel x86-based COTS server. Avoiding vendor lock-in can reduce the cost of deploying and upgrading the network, and can also lower the entry barriers for small vendors, bringing more innovation

into the market. Currently, MNOs are in different stages of adopting O-RAN, with more than 30 MNOs participating in the Open RAN project group in TIP. Open RAN is an evolution of virtualized RAN, as illustrated in Figure 8.

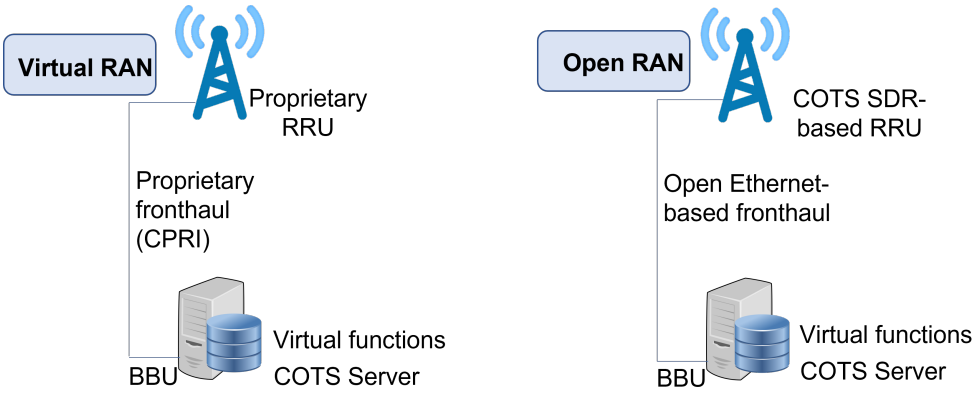

Figure 8. The two main changes in the evolution from virtual RAN to Open RAN are the commodity remote radio unit (RRU) and the open fronthaul (FH) interface. Note that 3GPP TR 38.801 has divided the baseband unit (BBU) into the centralised unit (CU) and the distributed unit (DU), where a single DU usually manages multiple RRUs. The common public radio fronthaul interface (CPRI) is a vendor-specific interface.

3.3.1. RAN Intelligent Controllers

The RAN Intelligent Controllers (RICs), including non-real-time (>1 s) and near-real-time (>10 ms and <1 s) operations, lie at the heart of the O-RAN architecture and introduce intelligent management of radio resources, self-organisation, and self-optimisation in the RAN by applying closed-control loops between the RICs and the RAN functions. Traditionally, the optimisation of the RAN is a vendor-specific problem, but introducing RICs clearly disaggregates the RAN from its optimisation process. Through open interfaces and RICs, the RAN functionality becomes available to operators, service providers and other parties, who can also contribute to RAN optimisation, using AI/ML.

The non-real-time RIC resides outside the RAN at the service management and orchestration (SMO) layer and is responsible for optimising in non-real-time the operation of the RAN, e.g., the orchestration of RAN slices, as well as enhancing network policies and boosting the E2E network performance. To do that, the controller receives real-time RAN data over the A1 interface (see Figure 9); data from other sources, such as the SMO framework, e.g., topology and configuration data; data from external repositories, e.g., contextual information; and historical data from the management data analytics function (MDAF). By applying big data analytics and AI/ML tools, the non-real-time RIC can optimise not only the performance at the RAN, but can also suggest joint network and RAN policies, e.g., coupling together the orchestration of network slices with radio resource scheduling techniques, making a step towards cross-layer network optimisation.

To give some representative examples of network optimisation using the non-real-time RIC, let us consider an emergency response scenario, where public safety personnel is deployed in the affected area. Using data collected from various sources, such as the map of the area received from an external database, the non-real-time RIC can prioritise the personnel operating at critical locations. Given the location of an UE, the non-real-time RIC may also apply traffic steering so that an emergency voice call is served by a macro cell, while eMBB video transmission from the scene to the command-and-control centre is let through a small cell in the area that operates at mm-wave frequencies. These sorts of functionalities are not supported by existing wireless networks; therefore, it is the non-real-time RIC that has mostly attracted the interest of the telecommunication industry.

In addition, the non-real-time RIC is responsible for ML model selection and training, while the near-real-time RIC carries out the ML inference once it receives the selected and

trained models. Recall from Section 3.2 that in the architectural framework proposed by ITU, an MNO can look at the ML marketplace for available models and use the ML sandbox to train and test them before deployment. At the specified timescales, the non-real-time RIC handles, for instance, with interference and spectrum management, traffic steering, handover, and resource allocation for specific users or slices. The ML models implementing these functions are embedded into the non-real-time RIC, in the form of programmable applications (xApps), which might belong to different vendors. The near- and non-real-time RIC are connected through the A1 interface but can also communicate via the O1 interface as illustrated in Figure 9. The non-real-time RIC controls through the E2 interface the centralised and distributed unit (CU/DU) functions. The radio unit (RU), DU, and CU are connected via the O1 interface to the RICs, periodically providing measurements and feedback. Additionally, the RU/DU/CU and the non-real-time RIC are managed by the SMO layer over the O1 interface.

Finally, the real-time closed-control loop (<10 ms) operates at the timescales less than the duration of the 5G new radio (NR) frame and can be used, for example, for adaptive modulation and coding, channel estimation, and channel prediction; see also Table 1. The real-time RIC is yet to be considered by the O-RAN Alliance. Clearly, different timescales are associated with different orders in the number of affected users [51].

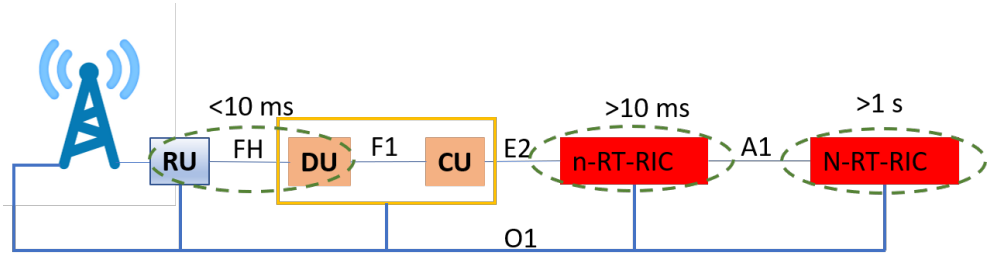

Figure 9. The open interfaces in the architecture proposed by the O-RAN Alliance and the three closed-control loops. Note that some of the functionality of the DU, e.g., low PHY layer operation, might move into the RU depending on the use case and deployment constraints and is known as split 7.2. Traditional 4G deployments use split 8. Moving functions from the DU to RU reduces latency and decreases the fronthaul bandwidth requirements at the cost of added complexity in the RU. See [99] for an overview of functional splits. The RU is deployed at the cell site, the CU/DU are usually located at the edge, while the RICs might be placed at the edge or cloud depending on the use case. One near-real-time RIC can control thousands of CUs, DUs, gNBs or eNBs.

3.3.2. Use cases and challenges in Open RANs

The O-RAN Alliance has already specified several use cases leveraging the O-RAN architecture and AI/ML. These use cases, along with their key benefits, are listed in Table 4 as potential research topics [100]. Of them, traffic steering was studied in [101], using a network architecture compliant to that proposed by O-RAN. Specifically, an LSTM is trained in the non-real-time RIC based on data collected from a real-world ultra-dense data traffic scenario. The trained NN is sent to the non-real-time RIC to predict the traffic in the considered area and proactively apply cell splitting by activating small cells inside the congested cells. Another use case, namely, scheduling control in sliced 5G networks through the non-real-time RIC, is investigated in [51]. A DRL agent selects the best scheduling policy, among water-filling, round-robin, proportional-fair, or max–min, for each base station and RAN slice. Both studies in [51,101] report promising results by adopting closed-control optimisation loops based on AI/ML in Open RANs.

The adoption of Open RAN architecture does not come without challenges. Firstly, there might be conflicts between xApps developed by different vendors. Consider, for instance, ML models produced independently for load balancing and handover that run in parallel. The former might allocate a UE to a neighbouring cell that is less congested,

and the latter might handover the UE back, creating ping-pong effects. As a result, it is incumbent to resolve inter-operability issues and xApps conflicts.

Secondly, Open RAN needs to work together with the existing wireless technologies that will still occupy a significant market share in 2025. In this direction, the Open RAN project group inside TIP focuses not only on 5G, but also on implementing software-defined 2G/3G/4G RANs with general-purpose hardware. Actually, most of today's Open RAN deployments are 4G. Thirdly, customised Open RAN solutions are needed for private 5G deployments with specialised use cases. Finally, since the network elements can belong to different vendors, suboptimal performance should be expected in early developments. Therefore, it will not be possible to tell whether Open RAN is the leading choice until we witness deployments in dense urban scenarios, with integrated massive MIMO and not just Open RAN in rural environments.

Table 4. Use cases considered by the O-RAN alliance [100].

Use Cases	Key Benefits and Enablers
Low-cost RAN White-box Hardware	COTS hardware reduces CapEx. It is easier to upgrade and inter-operable with network functions developed by different vendors.
Traffic Steering	Faster response to data traffic variations using AI/ML-based proactive load-balancing yielding reduced OpEx, better network efficiency and user experience.
QoE Optimisation	Prediction of degraded QoE for a UE using AI/ML and proactive allocation of radio resources to the UE.
QoS-based Resource Optimisation	AI/ML-based allocation of radio resources to ensure that at least certain prioritised users maintain their QoS under data traffic congestion.
Massive MIMO Optimisation	Adapting beam configuration and related policies, e.g., packet scheduling, for enhancing the network capacity.
RAN slice SLA assurance	Maximise revenue with AI/ML-based management of network slices.
Context-based dynamic handover management for vehicle-to-everything (V2X)	AI/ML-based handovers using historical road traffic and navigation data resulting to better user-experience.
Dynamic resource allocation based on the flight-path for unmanned aerial vehicles (UAV)	AI/ML-based resource allocation using historical flight data and UAV measurement reports.
Radio resource allocation for UAV applications	AI/ML-based resource allocation under asymmetric uplink/downlink data traffic.
RAN sharing	Reduced CapEx due to multi-vendor deployments.

The O-RAN Alliance created the O-RAN software community inside the Linux foundation to professionally manage its open-source software. Since November 2019, there have been three software releases, with the fourth planned for July 2021. Rakuten was the first to implement a multi-vendor RAN in 2019. In 2020, NTT Docomo successfully connected a 5G baseband unit developed by NEC and Samsung with RRUs from other vendors complying with O-RAN specifications. In the U.K., Vodafone has already deployed rural Open RAN sites; see [102] for more case studies. Nevertheless, O-RAN deployments have so far concentrated on vendor interoperability; implementing intelligent controls (the RICs), which are key pieces in the O-RAN architecture, is not expected to run commercially until the end of 2021.

3.4. TinyML

TinyML lies in the intersection of ML and embedded systems, and enables ML capabilities on lightweight IoT devices, including cheap Micro-controller units (MCUs) or digital signal processors, which operate at an average power of 1 mW or less [103]. These devices (without the ML functionality yet) have become an integral part of our everyday lives, being massively deployed at home appliances and buildings, mobile phones, and automotive, to name a few. During 2019, more than 30 billion MCUs were shipped worldwide.

MCU-based IoT lies at the very edge of the communication network, collecting enormous amounts of raw data, most of which is typically discarded, as the energy cost to transfer it to the cloud is prohibitively high. Note that for devices falling under the category of TinyML, the energy consumption required for communication is much higher than that needed for computation. Keeping the total energy consumption at very low levels is actually a prerequisite for the massive deployment of MCU-based devices, as these devices can be left unattended for months or years after their deployment, operating on battery, solar energy, or perhaps energy harvesting. Therefore, empowering MCUs with AI/ML and doing locally ML inference while the overall power consumption remains low can be beneficial. For instance, a motion sensor often calculates and transmits to a gateway just its peak and average acceleration values every duty cycle. Now, consider that the sensor runs a lightweight version of an ML model, which constantly monitors the device and can predict a failure. The sensor will communicate to the cloud just a few extra bits of information if a potential failure is predicted. Environmental and agricultural monitoring, anomaly detection, and predictive maintenance are relevant use cases for TinyML.

The main limitation of TinyML is the hardware constraints of embedded systems in terms of processing power, memory, and clock speed. Typically, TinyML runs on 32 bit MCUs with fewer than 500 kB of RAM, a few MB of flash memory, and a clock frequency lower than 200 MHz [104]. Representative examples are the Sparkfun edge Arduino board, the ARM Cortex-M7 MCU with 320 kB of RAM and 1 MB of storage, and the QCC112 Chipset by Qualcomm for ultra-low power always-on computer vision [105]. Notably, TinyML does not include end-devices carrying micro-computers, such as Rasberry Pi, which can run Linux, support high-level programming languages, such as Python, and typically has 2–3 orders of magnitude more memory than MCUs. Due to the limited available resources, TinyML shall be used only for ML inference while training is executed off-board and offline. Additionally, the ML task should be specific, which usually translates to a binary output layer, e.g., whether a time series is abnormal or not, and keyword spotting. The authors of [106] have designed a NN on off-the-shelf MCUs and accelerated the inference of wake word applications by three times.

Recent advances in ML algorithms running at the edge and cloud do not apply to tinyML, where the size of the NNs has to be limited. Even though the training and validation are executed at the cloud, using traditional libraries, such as Pytorch and Tensorflow (TF), the ML model would be probably too large to store in embedded systems. DNNs running on MCUs should be optimised to adapt to the power, memory, and processing constraints of the targeted hardware. In this direction, the TF Lite Micro is an open-source interpreter developed by Google that figures out how to compress pre-trained ML models generated in traditional frameworks, such as TF, to a size suitable for TinyML [107]. The STM32Cube.AI is another tool that automatically converts pre-trained artificial neural network (ANN) models to pieces of code that can run on MCUs, providing also feedback about the performance on the selected hardware. While customised codes can perform better than ML interpreters by optimally reducing the data dimensionality, or quantising and pruning the trained model, the price paid is the reduced portability to other devices [108].

To conclude, TinyML is a fast-growing field but more advances in hardware, software, benchmarking, ML algorithms, and applications are required to see omnipresent TinyML devices [109].

4. Overview of Research Projects on AI/ML for Communications and Networking

In this section, we review the major outcomes of collaborative research projects regarding intelligent wireless networks. Most of the consortia comprise a mix of research institutions and telecommunication industry, involving vendors, MNOs, small and medium enterprises (SMEs), and communication service providers (CSPs). We emphasise on the selected use cases and associated research problems that can be derived from each project, and highlight how AI/ML tools help satisfy the respective technical requirements. We start our review with European research grants, where more reports and presentation slide sets are publicly available, followed by a summary of the activities of research frameworks in the U.K. and the U.S.

4.1. Europe—H2020 Research Framework

The recently published European vision for the 6G ecosystem suggests an "AI everywhere" principle, where AI is infused throughout the network, can be traded as a service, and is ultimately the key enabler to intertwine the human, physical and digital worlds. AI shall eventually enable a self-contained wireless ecosystem with zero-touch and automated decision-making processes, where massive data collection, distributed computation, and ML inference inter-work to meet stringent network performance targets and offer an immersive experience to subscribers [110].

The EU Horizon 2020 program has already put a strong emphasis on the research of AI/ML-empowered wireless networks [111]. In the latest call 5G-PPP Phase-3, there are in total nine topics, with most of the funded projects scrutinising AI/ML-based solutions for communications and networking in their objectives [112]. The nine project themes are as follows: coordination and support actions, infrastructure projects (5GENESIS), advanced 5G validation trials across multiple vertical industries (5GROWTH), long-term evolution (ARIADNE), automotive (5G-CARMEN), 5G core technologies innovation projects, 5G for connected and automated mobility, innovation for verticals with third-party services, and smart connectivity beyond 5G (the flagship Hexa-X project). Inside the parentheses, we include representative projects from each theme that are elaborated shortly below.

In a nutshell, most of the grants under the 5G-PPP Phase-3 call apply AI/ML for dynamic network slicing and closed-loop automation at the edge and core networks. Furthermore, the near-real-time RIC and the software-based control of reflective intelligent surfaces (RIS) constitute the lion's share of research activities in the RAN. More research outputs regarding the real-time RIC are available under the Windmill project within the EU H2020 Marie Skłodowska-Curie framework innovative training networks (ITN) [113].

4.1.1. ARIADNE

Artificial intelligence aided D-band network for 5G long-term evolution (ARIADNE) is a three-year research and innovation project that started in November 2019 and is funded with EUR 6 million under the EU framework, Horizon 2020 (H2020). The overarching goal of the project is to investigate techniques and technologies that can efficiently support high-bandwidth communications in the D-band (130–174.8 GHz). This band is a promising spectrum candidate for implementing the fronthaul and backhaul of small cell networks operating in mm-wave frequencies and for enabling very high data rates over short-range connections in the RAN [114]. Three use cases and six corresponding scenarios are selected by the consortium and are summarised in Table 5 below [115]. Concerning the selected scenarios, the project sets the following KPI targets: an aggregate throughput (or an E2E throughput where applicable) of 100 Gbps, nearly 100% connectivity reliability, and 10 times energy efficiency improvement in the RAN, as compared to public 5G RAN deployments.

Table 5. Use cases and corresponding scenarios considered by ARIADNE.

Use Cases	Scenarios	Mobility
Outdoor backhaul and fronthaul networks of fixed topology.	Long-range line-of-sight (LoS) rooftop point-to-point backhauling without RIS.	Stationary
	Street-level point-to-point and point-to-multipoint backhauling and fronthauling with RIS for non-LoS (NLoS).	Stationary
Advanced NLoS connectivity based on meta-surfaces.	Indoor advanced NLoS connectivity based on meta-surfaces.	Stationary or low.
	Data-kiosk communication for downloading large amounts of data in a short time.	Stationary or low or moderate.
Ad hoc connectivity in moving network topology.	Dynamic fronthaul and backhaul connectivity for mobile 5G access nodes and repeaters, e.g., using drones.	Low or moderate.
	LoS vehicle-to-vehicle (V2V) and vehicle-to-everything (V2X) connectivity.	Low or moderate.

The opportunities brought forward by AI/ML to achieve the objectives of ARIADNE are vast. For instance, in the D-band, highly directional antennas should be used to overcome the adverse signal propagation conditions. However, with very narrow beams, the capacity of a long-range (<500 m) backhaul link becomes sensitive, even to small beam misalignments (due to winds or other physical phenomena), calling for continuous channel estimation to correct the beam steering. In short-range scenarios (10–30 m) involving mobility, accurate user localization for beam tracking is required. Therefore, ML techniques for real-time channel estimation and prediction (see Table 1), and ML-based mobility and beamforming prediction become relevant [116]. For NLoS environments, implementing highly reliable and high data rate backhaul and fronthaul with RIS imposes many challenges, especially on the optimal design of phase shifts (see Table 1). While the focus of ARIADNE is the real-time RIC (timescales of less than 10 ms), several near-real-time closed-control functions, such as load prediction at the backhaul, power control, and user scheduling based on AI/ML are also investigated. Finally, channel modelling in the largely unexplored D-band, including, for example, models for the link gain and the channel impulse response under various weather conditions and user mobility, is another potential area for applying AI/ML techniques.

4.1.2. 5GENESIS

The 5th generation E2E network, experimentation, system integration, and showcasing (5GENESIS) is a 30-month 5G PPP project that started in July 2018 and received approximately EUR 16 million in contributions from the EU. The experimental facilities of 5GENESIS include five testbeds that are used to implement several 5G-enabled use cases; see Table 6. To associate the measured KPIs with the network status, monitoring techniques and probes are implemented at several locations, network components and layers. In particular, data are collected at the end devices to measure the radio conditions and the device power consumption, at the RAN to evaluate resource availability and track the traffic load at the backhaul and fronthaul, at the core network to compute the processing load of various NFs, and at the edge/cloud to measure the CPU consumption and the utilisation of available computational resources. Then, ML is employed to process the large volumes of extracted data and identify hidden associations and correlations between the measured KPIs and the monitored network parameters and conditions. The measured correlations and causalities could be used at the real-time RIC, e.g., to design intelligent adaptive modulation and coding schemes, or may also lead to prescriptive analytics, e.g., making suggestions about how much additional computing power and storage should be deployed at the edge to meet a target KPI.

Table 6. Use cases associated with KPIs and 5G service types considered by 5GENESIS [117].

Use Cases	Latency	Coverage	Service Creation Time	Capacity	Availability	Reliability	User Density	Service Types
Big event	✓	✓	✓					eMBB
Eye in the sky	✓	✓	✓					eMBB, uRLLC
Security as a service at the edge	✓		✓					All
Wireless video in large scale event			✓	✓	✓			eMMB
Multimedia mission critical services	✓			✓		✓	✓	eMBB, uRLLC
MEC-based mission critical services	✓		✓	✓	✓			eMMB
Maritime communications	✓	✓		✓		✓		eMMB
Capacity on demand and rural IoT	✓			✓		✓		eMMB, mMTC
Massive IoT for large-scale public events	✓		✓	✓		✓		mMTC
Dense urban 360 degrees virtual reality	✓		✓	✓		✓		eMBB

Let us consider the use case "eye in the sky" of 5GENESIS where an UAV is controlled over 5G, where both the vehicle and the pilot are being served bydifferent cells. In that case, an uRLLC slice is required to route the joystick signals controlling the drone manoeuvres through the 5G core, and prioritise them over other applications with less stringent latency requirements. At the same time, an eMBB-with-controlled-latency slice must be also prioritised to route the real-time 4K video footage collected from the drone to the ground near the pilot through the 5G core [117]. In such a complex scenario, AI/ML could uncover how slices at the RAN and the core network could be stitched together, given the radio propagation conditions and the traffic congestion at the core network under an E2E latency constraint. The This use case demonstrates the use of ML for E2E network multi-slice operations.

4.1.3. 5GROWTH

5G-enabled growth in vertical industries (5GROWTH) is a 33-month EU-funded project (approximately EUR 14 million) that started in June 2019 with the vision to design an AI-based 5G E2E solution for vertical industries to meet their respective performance targets, simultaneously. The verticals involved in the project, including Industry 4.0 and transportation, have specified use cases that have very diverse technical requirements. For instance, Industry 4.0 COMAU, a world leader in industrial automation, opts for digital twins, smart factory telemetry and monitoring apps, digital tutorials, and remote support, while in the transportation industry, EFACEC is likely to focus on safety communications [118]. On the one hand, the technical requirements of telemetry applications are high availability 99.999%, low latency 15–100 ms, low-mobility support 3–50 km/h, and high connection density ≈5000 devices/km^2. On the other hand, a use case in railway transportation, such as real-time, high-definition image transmission from a surveillance camera at the level crossing to the train approaching the intersection, requires sufficient macrocellular coverage and eMBB support at high speeds ≈160 km/h [118]. To accommodate such diversity of technical requirements over the same physical infrastructure, a sophisticated management of network slices is needed, which opens new research problems, where AI/ML can play a central role.

For an efficient use of the created network slices, the concept of slice-sharing among multiple (vertical) services can be applied. Furthermore, a single service can be decomposed into sub-services with different functionalities, such as the Services C and D in Figure 10, which are served from different slices, leveraging the idea of network slice subnet instance (NSSI) [119]. Given the complexity of the selected use cases, it is not surprising that the consortium has contrived four innovative technology components for efficient network slicing using AI/ML [120]. These components implement dynamic and intelligent management of network slices and integrate AI/ML throughout the network for optimal orchestration and resource control, anomaly detection, QoS forecasting, and inference.

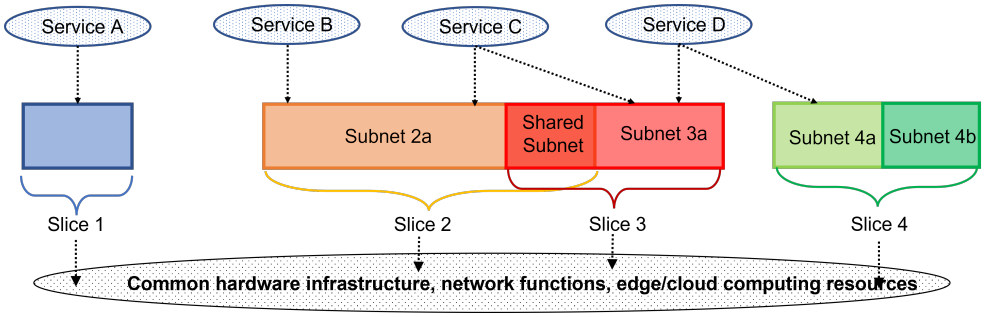

Figure 10. Service A exclusively uses Slice 1. Leveraging the principle of NSSI, Slices 2, 3 and 4 are broken down into two subnets. Slice 2 and Slice 3 share a subnet. Services C and D are decomposed into two sub-services and one of them is served by the Subnet Slice 3a. For instance, Subnet 2a and Subnet 4a might be core network slices, while Subnet 3a might be a RAN slice. In this illustration, the Services C and D need similar functionalities at the RAN, but not at the core network; hence, they can share a RAN NSSI but use different core network NSSI. In this way, the created slices are used more efficiently as compared to the exclusive allocations of network slice instances to network services.

4.1.4. 5G-CARMEN

5G for connected and automated road mobility in the European Union (5G-CARMEN) (approximately EUR 19 million EU contribution between November 2018 and October 2021) focuses on four automotive use-case families, namely, cooperative manoeuvring, situation awareness, video streaming, and green driving. Big data analytics and AI/ML tools become essential in addressing the KPIs associated with the selected use cases [121].

Firstly, the cooperative manoeuvring and situation awareness use cases comprise several subcases such as cooperative lane merging, back-situation awareness of an emergency vehicle arrival, and motorbike awareness. To improve safety and enhance the ride quality at complex merges, vehicles share their current position, speed, and trajectory with a roadside unit (RSU). A MEC server at the RSU processes the received data, using AI/ML, and recommends optimal actions for the drivers, e.g., speed profiles creating sufficient gaps between successive vehicles. As a result, accurate position information, in the order of a meter or less, also becomes essential [122]. In this regard, cooperative perception algorithms at the vehicles that optimally fuse the location data received over the global navigation satellite system and inertial monitoring unit (GNSS/IMU) with that obtained from on-board sensors, such as lidar and cameras, can be employed to meet the required constraints on localisation accuracy before the vehicles share their locations with the RSU. Furthermore, in the considered use cases, the exchange of communication messages between the vehicles and the RSU is likely to involve short packets to satisfy the low-latency communication constraints, and thus, ML-based decoding of channel codes also becomes relevant; see Table 1.

Secondly, the video streaming use case concerns the continuous reception of high-definition on-demand video while onboard and in challenging situations, where the video quality is likely to deteriorate, due to handovers in cross-border areas. In this scenario, given the trajectory of the subject vehicle, AI/ML models running at the edge/cloud can assist in predicting the quality of the received video stream, while driving through the cross-border area. To do that, the 5G-Carmen platform should account for the historical data about the propagation channel conditions in the area, the current demand at the RAN, and the traffic congestion at the core network. Besides predictive analytics, the AI/ML algorithms shall also prescribe in advance appropriate courses of action to avoid streaming disruptions, e.g., prompt the downloading of more video chunks when the network conditions are strong to compensate while travelling across the problematic areas [122].

Thirdly, the objective of green driving is to improve the environmental quality in sensitive regions by regulating the density of road traffic and the emissions of hybrid vehicles with an electric powertrain. This use case involves extensive data collection from smart sensors on the vehicles and RSUs for monitoring the air quality, which is then sent to the edge/cloud for analysis and inference. Using AI/ML, the collected data are associated with historical measurements, the current weather conditions, and the types of vehicles on the move to showcase the air quality status and its predicted future quality to the responsible stakeholders. Apart from showcasing, the 5G-CARMEN system ultimately supports prescriptive analytics suggesting environment-friendly actions, e.g., speed profiles for individual vehicles depending on their type (motorcycles, cars, or lorries), alternative routes for non-electric vehicles, and definition of electric zones to relieve air pollution in critical areas [122].

4.1.5. Smart Connectivity beyond 5G

Under this topic, there are nine funded projects, all started in January 2021 with a duration between 24 and 42 months that have received an aggregate EU contribution of approximately EUR 60 million. All projects make extensive use of AI/ML tools in their respective target areas as summarised in Table 7. The AI/ML aspects of the Hexa-X project, a flagship for B5G/6G vision and intelligent fabric of technology enablers connecting human, physical, and digital worlds are discussed next.

Table 7. Funded projects under the call ICT-52-2020—5G PPP—Smart Connectivity beyond 5G, along with their main application areas of AI/ML and target use cases.

Project Name	Target Areas Using AI/ML
6G Brains	Implements a self-learning agent based on deep reinforcement learning which performs intelligent and dynamic resource allocation for future industrial IoT at massive scales.
AI@Edge	Designs reusable, trustworthy and secure AI solutions at the network edge for autonomous decision making and E2E quality assurance. The targeted use cases are AI-based smart content pre-selection for in-flight infotainment, AI-assisted edge computing for infrastructure monitoring using drones, and AI for intrusion detection in industrial IoT.
Daemon	Implements AI/ML algorithms for real-time network control and intelligent orchestration and management. Specifically, the project investigates the use of real time RIC for embedding intelligence in RIS, radio resource allocation, distribution of computational resources at the edge/fog, and backhaul traffic control. At longer timescales, energy-aware network slicing, capacity forecasting, anomaly detection and self-learning network orchestration are examined.
Dedicat 6G	Aspires to demonstrate distributed network intelligence for dynamic coverage extension, indoor positioning, data caching, and energy-efficient distribution of computation loads across the network. Representative use cases demonstrate the developed solutions include smart warehousing, augmented and virtual reality applications, public safety and disaster relief using automated guided vehicles and drones, and connected autonomous mobility.
Hexa-X	Applies AI for network orchestration from the end-devices through the edge to the cloud and the core network. This includes inter-connecting intelligent agents based on federated learning, proactive network slice management, instantiation of network functions, zero-touch automation, explainable AI, intelligent spectrum usage and intelligent air interface design, to name a few.
Marsal	Integrates blockchain technology with ML-based mechanisms to foster privacy and security in multi-tenant network slicing scenarios. Furthermore, ML-based orchestration and management of radio and computational resources is studied.
Reindeer	Develops an experimental testbed for the RadioWeaves technology. RadioWeaves leverages the ideas of RIS and cell free wireless access for offering zero-latency and high-capacity connectivity in short-range indoor applications such as immersive entertainment, health care, and smart factories. Intelligence is distributed near the end devices for an efficient use of spectral, energy and computational resources.
Rise-6G	Designs and prototypes intelligent radio-wave propagation using RIS.
Teraflow	Implements a novel SDN controller with cloud-native architecture, AI-based security, and zero-touch automation features. The demonstrated use cases are autonomous networks beyond 5G, cybersecurity and automotive.

The vision of Hexa-X is to intertwine the human, physical and digital worlds for sustainable, trustworthy, and inclusive use of communication technologies. In this direction, the project has, so far, identified five use-case families and associated them with six research challenges. On the one hand, the use cases are sustainable development, massive twinning, immersive telepresence for enhanced interaction, from robots to cobots (robots developing relations among each other), and local trust zones for humans and machines. On the other hand, the research challenges are connecting intelligence, extreme experiences, a network of networks, sustainability, trustworthiness, and global service coverage. According to [123] (Figure 4.1), connecting intelligence is required to address the first four of the above-mentioned use-case families.

In addition to the identified use cases, scenarios and challenges, the project has also specified seven services harnessing new capabilities, including AI-as-a-service and AI-assisted V2X. The former applies the producer–consumer model between the end-devices and third-party AI agents that offer trained ML models for inferencing tasks. The latter generates digital twins of urban cities and adopts AI techniques to regulate the

road traffic by recommending actions that are communicated over V2X links back to the vehicles. Finally, Hexa-X considers aspects of spectrum evolution and studies intelligent spectrum management and interference mitigation techniques. For instance, in industrial IoT environments, radio usage often experiences periodic cycles, and thus, AI-based interference prediction schemes could be used to avoid interference in certain times and areas and, subsequently, enhance the spectrum utilisation efficiency.

4.2. U.K.-EPSRC and U.S.-NSF

The Engineering and Physical Sciences Research Council (EPSRC) in the U.K. has already funded 250 projects on AI technologies with an aggregate budget of around GBP 178 million, which roughly covers 3.2% of EPSRC's total investment portfolio [124]. Out of these grants, 100 projects fall under information and communication technologies (ICT), 24 grants are engineering focused, and 18 grants are used to train doctoral students. However, only a limited number of projects concern the integration of AI/ML into the B5G/6G wireless ecosystem.

The LEANCOM project (GBP 0.9 million, November 2019 to November 2022) combines existing mathematical models for the performance evaluation at the PHY layer with deep learning, with an emphasis onlow-cost devices with hardware imperfections,which are overlooked by existing mathematical models [125]; recall from Section 2 the discussion on the model-aided ML. The 6GRADIO project (GBP 0.5 million, December 2020 to November 2023) equips the Mitola radio [126] with the power of collective intelligence, using an intelligence gathering mechanism leveraging both game-theoretic approaches and DRL. Next, the overarching goal of the SWAN project (GBP 2.3 million, February 2020 to January 2025) is network security against cyber-attacks and resilience to network faults and failures. In this direction, SWAN applies AI/ML under four main pillars, namely, threat synthesis and assessment to proactively identify vulnerable interfaces, radio-frequency detection of cyber-attacks for risk mitigation, cyber-secure and agile design of waveforms using software-defined radios, and secure dynamic spectrum access [127]. Finally, the CONNECT project, funded by EPSRC through the European CHIST-ERA framework, is about collaborative and intelligent decision making at the edge through novel data caching, distributed computing, and federated learning [128], see Section 3.1.1 for a discussion on federated learning.

Besides standard research projects, EPSRC has also created program grants, which bring together world-leading researchers to address significant challenges. One example in the fields of AI and networking is the TRANSNET grant (GBP 6.1 million, August 2018 to July 2024), intending to create an intelligent optical network, i.e., the backbone of the future digital infrastructure, which can dynamically allocate resources for maximising the network performance, where and when needed [129]. Under the same framework, EPSRC also funds theoretically oriented consortia, e.g., the MATHDL project (GBP 3.4 million, September 2021 to August 2026) aspires to develop a mathematically rigorous foundation for the training, inference, and performance of deep learning to enhance its explainability and trustworthiness [130]. Similarly, the GRAPHNEX project, funded through CHIST-ERA, focuses on explainable AI by decomposing highly connected DNNs into smaller units, which perform specific tasks with interpretable outcomes [131].

In the U.S., a partnership between Intel and the National Science Foundation (NSF) on ML for wireless networking systems was formed in 2019 [132]. The initiative has two main pillars, i.e., ML for spectrum management and distributed ML over wireless edge networks [133]. In a nutshell, under the first pillar, the funded projects cover model-based ML at the PHY layer to address poorly understood non-linear distortions and relax simplified assumptions on impairment models, construct robust channel codes under interference and enhance the network performance through intelligent power control, beamforming, scheduling, distributed spectrum sensing and spectrum access. Under the second pillar, most projects investigate how to realise collective intelligence at the edge, using enhanced and multi-hop federated learning.

In April 2021, NSF announced a new partnership with private industries on intelligent and resilient next-generation systems with an anticipated funding of approximately USD 40 million, which seeks to accelerate the adoption of intelligent decision making in future networking systems as a means to enhance their resilience under extreme operating conditions [134]. At least two of the primary considerations of this call are relevant to the use of AI/ML techniques in next-generation networks [135]. Firstly, AI can be used to predict disruptions in the availability of resources and prescribe solutions to sustain an acceptable quality of the offered service, e.g., in V2X scenarios, historical data and radio environmental maps can be used to predict the degradation in signal reception quality while travelling through some problematic areas. Then, AI algorithms may trigger beforehand, when the channel quality is good, the downloading of more data packets for graceful service performance degradation. Secondly, investigating the required resilience in computational capabilities for sustaining a good-enough performance in distributed learning schemes under faults and failures is another important topic to pursue. For example, let us consider federated learning, where some local learners fail to send their model updates to the server, i.e., a UE runs out of battery. Developing an AI-based component that can predict these failures and advise appropriate countermeasures to mitigate the performance loss is a rather unexplored topic.

5. Conclusions

The ongoing deployments of 5G wireless networks mainly support eMBB services, but it is envisaged that 5G/6G networks would penetrate across various vertical industries and offer many more customised services and applications. On the one hand, verticals have already identified several profitable use cases that make use of the unprecedented high data rates low latencies and high numberof connected devices offered by emerging networks. On the other hand, communication service providers have identified that AI/ML-based network orchestration is the only way to efficiently and dynamically support stringent and diversified technical requirements of various use cases over the same unified physical infrastructure. Note that deploying dedicated networks to each vertical is a wasteful usage of scarce and limited resources such as frequency spectrum and core network hardware components. Also, reactive hand-engineered policies and static thresholding rules for network slicing, resource allocation and RAN management are not tailored to heterogeneous networks and diversified use cases resulting in sub-optimal operation and inefficient use of the available resources.

Our literature review reveals that network monitoring, extensive data collection, and AI/ML will become the main tools to optimise, manage and orchestrate current and future networks. Ongoing work in 5G standardisation includes the network data analytics function (NWDAF) and the management data analytics function (MDAF) developed by 3GPP, and the RAN intelligent control (RIC) recommended by the O-RAN Alliance for coordinating the data collection, training, and inference of ML models at the core and the RAN, respectively. We have found in the existing literature several studies on real-time RIC and federated learning at the core network. These studies are important preliminary steps towards RAN control at longer time scales and distributed connected intelligence at the edge and the core network, which can enhance the performance in a multitude of use cases. Innovations in data science and communication technologies could evolve independentlyand be disentangled from one another, as suggested by ITU-T FG-ML5G, using digital twins and the concept of ML sandbox to compareML models before deployment. This will ease the integration of newly developed AI/ML models into future networks. Finally, closed-loop controls , developed by the ETSI ZSM committee, will go beyond local network processes, e.g., congestion control at the transport layer, and we think they will become the prevailing solution for zero-touch network automation.

In this survey, we have also reviewed the major and most impactful large-scale research projects towards intelligent communications & networking. Intelligent real-time RIC, e.g., intelligent channel coding, can bring significant benefits to network operators [136], and help address significant challenges such as urban V2X communication using software-controlled reflecting surfaces. Nevertheless, the focus of the telecommunication industry

has been so far data to collect data at the RAN and the core network for intelligent control at timescales larger than 1000 ms. Apart from network management and orchestration, other representative examples at this timescale include AI/ML-based E2E network slicing, dynamic content selection for edge caching, and distributed learning such as forming federations between several NWDAF entities. To conclude, , we believe that, ultimately, AI/ML would become the lifeblood of future networksand be vital in shaping the orchestration of available technologies, bandwidth, and computational resources to meet the customer intents, given energy and cost constraints.

Author Contributions: Supervision, M.H., J.E., M.I. and R.T.; Writing—original draft, K.K.; Writing—review & editing, K.H. and M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are often used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
AWGN	Additive White Gaussian Noise
CPU	Central Processing Unit
CQI	Channel Quality Indicator
CSI	Channel State Information
DL	Deep Learning
DNN	Deep Neural Network
DRL	Deep Reinforcement Learning
eMBB	Enhanced Mobile Broadband
E2E	End-to-End
KPI	Key Performance Indicator
LSTM	Long Short Term Memory
MCU	Micro-Controller Unit
MDAF	Management Data Analytics Function
MEC	Multi-access Edge Computing
ML	Machine Learning
MNO	Mobile Network Operator
NFV	Network Function Virtualisation
NN	Neural Network
NDDI	Network Slice Subnet Instance
NWDAF	Network Data Analytics Function
QoE	Quality-of-Experience
RAN	Radio Access Network
RIC	RAN Intelligent Control
RIS	Reflecting Intelligent Surfaces
RNN	Recurrent Neural Network
SDN	Software Defined Networking
UE	User Equipment

References

1. Li, R.; Zhao, Z.; Zhou, X.; Ding, G.; Chen, Y.; Wang, Z.; Zhang, H. Intelligent 5G: When Cellular Networks Meet Artificial Intelligence. *IEEE Wirel. Commun.* **2017**, *24*, 175–183. doi:10.1109/MWC.2017.1600304WC.
2. Letaief, K.B.; Chen, W.; Shi, Y.; Zhang, J.; Zhang, Y.J.A. The Roadmap to 6G: AI Empowered Wireless Networks. *IEEE Commun. Mag.* **2019**, *57*, 84–90. doi:10.1109/MCOM.2019.1900271.
3. Jiang, W.; Han, B.; Habibi, M.A.; Schotten, H.D. The Road Towards 6G: A Comprehensive Survey. *IEEE Open J. Commun. Soc.* **2021**, *2*, 334–366. doi:10.1109/OJCOMS.2021.3057679.

4. Zhao, Y.; Zhai, W.; Zhao, J.; Zhang, T.H.; Sun, S.; Niyato, D.; Lam, K.Y. A Comprehensive Survey of 6G Wireless Communications. *arXiv* **2020**, arXiv:2101.03889.
5. Bhat, J.R.; Alqahtani, S.A. 6G Ecosystem: Current Status and Future Perspective. *IEEE Access* **2021**, *9*, 43134–43167. doi:10.1109/ACCESS.2021.3054833.
6. Hossain, M.A.; Noor, R.M.; Yau, K.L.A.; Azzuhri, S.R.; Zaba, M.R.; Ahmedy, I. Comprehensive Survey of Machine Learning Approaches in Cognitive Radio-Based Vehicular Ad Hoc Networks. *IEEE Access* **2020**, *8*, 78054–78108. doi:10.1109/ACCESS.2020.2989870.
7. Praveen Kumar, D.; Amgoth, T.; Annavarapu, C.S.R. Machine learning algorithms for wireless sensor networks: A survey. *Inform. Fusion* **2019**, *49*, 1–25. doi:10.1016/j.inffus.2018.09.013.
8. Mahdavinejad, M.S.; Rezvan, M.; Barekatain, M.; Adibi, P.; Barnaghi, P.; Sheth, A.P. Machine learning for internet of things data analysis: A survey. *Digit. Commun. Netw.* **2018**, *4*, 161–175. doi:10.1016/j.dcan.2017.10.002.
9. Khalil, R.A.; Saeed, N.; Masood, M.; Fard, Y.M.; Alouini, M.S.; Al-Naffouri, T.Y. Deep Learning in the Industrial Internet of Things: Potentials, Challenges, and Emerging Applications. *IEEE Internet Things J.* **2021**, *8*, 11016–11040. doi:10.1109/JIOT.2021.3051414.
10. Mao, Q.; Hu, F.; Hao, Q. Deep Learning for Intelligent Wireless Networks: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 2595–2621. doi:10.1109/COMST.2018.2846401.
11. Zhang, C.; Ueng, Y.L.; Studer, C.; Burg, A. Artificial Intelligence for 5G and Beyond 5G: Implementations, Algorithms, and Optimizations. *IEEE J. Emerg. Sel. Top. Circ. Syst.* **2020**, *10*, 149–163. doi:10.1109/JETCAS.2020.3000103.
12. Ali, S.; Saad, W.; Steinbach, D.E. *White Paper on Machine Learning in 6G Wireless Communication Networks*; 6G Research Visions, No. 7; University of Oulu: Oulu, Finland, 2020.
13. Gündüz, D.; de Kerret, P.; Sidiropoulos, N.D.; Gesbert, D.; Murthy, C.R.; van der Schaar, M. Machine Learning in the Air. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2184–2199. doi:10.1109/JSAC.2019.2933969.
14. Sun, Y.; Peng, M.; Zhou, Y.; Huang, Y.; Mao, S. Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues. *IEEE Commun. Surveys Tutor.* **2019**, *21*, 3072–3108.
15. Boutaba, R.; Salahuddin, M.A.; Limam, N.; Ayoubi, S.; Shahriar, N.; Estrada-Solano, F.; Caicedo, O.M. A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *J. Internet Ser. Appl.* **2018**, *9*. doi:10.1186/s13174-018-0087-2.
16. Shafin, R.; Liu, L.; Chandrasekhar, V.; Chen, H.; Reed, J.; Zhang, J.C. Artificial Intelligence-Enabled Cellular Networks: A Critical Path to Beyond-5G and 6G. *IEEE Wirel. Commun.* **2020**, *27*, 212–217. doi:10.1109/MWC.001.1900323.
17. Zhang, C.; Patras, P.; Haddadi, H. Deep Learning in Mobile and Wireless Networking: A Survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 2224–2287. doi:10.1109/COMST.2019.2904897.
18. Wang, J.; Jiang, C.; Zhang, H.; Ren, Y.; Chen, K.C.; Hanzo, L. Thirty Years of Machine Learning: The Road to Pareto-Optimal Next-Generation Wireless Networks. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1472–1514.
19. Zappone, A.; Di Renzo, M.; Debbah, M.; Lam, T.T.; Qian, X. Model-Aided Wireless Artificial Intelligence: Embedding Expert Knowledge in Deep Neural Networks for Wireless System Optimization. *IEEE Veh. Technol. Mag.* **2019**, *14*, 60–69. doi:10.1109/MVT.2019.2921627.
20. Wang, T.; Wen, C.K.; Wang, H.; Gao, F.; Jiang, T.; Jin, S. Deep learning for wireless physical layer: Opportunities and challenges. *China Commun.* **2017**, *14*, 92–111. doi:10.1109/CC.2017.8233654.
21. Qin, Z.; Ye, H.; Li, G.Y.; Juang, B.H.F. Deep Learning in Physical Layer Communications. *IEEE Wirel. Commun.* **2019**, *26*, 93–99. doi:10.1109/MWC.2019.1800601.
22. O'Shea, T.; Hoydis, J. An Introduction to Deep Learning for the Physical Layer. *IEEE Trans. Cogn. Commun. Netw.* **2017**, *3*, 563–575. doi:10.1109/TCCN.2017.2758370.
23. O'Shea, T.J.; Karra, K.; Clancy, T.C. Learning to communicate: Channel auto-encoders, domain specific regularizers, and attention. In Proceedings of the 2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Limassol, Cyprus, 12–14 December 2016; pp. 223–228. doi:10.1109/ISSPIT.2016.7886039.
24. Balevi, E.; Andrews, J.G. One-Bit OFDM Receivers via Deep Learning. *IEEE Trans. Commun.* **2019**, *67*, 4326–4336. doi:10.1109/TCOMM.2019.2903811.
25. Mulvey, D.; Foh, C.H.; Imran, M.A.; Tafazolli, R. Improved Neural Network Transparency for Cell Degradation Detection Using Explanatory Model. In Proceedings of the ICC 2020—2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 7–11 June 2020; pp. 1–6. doi:10.1109/ICC40277.2020.9149391.
26. Shlezinger, N.; Eldar, Y.C.; Farsad, N.; Goldsmith, A.J. ViterbiNet: Symbol Detection Using a Deep Learning Based Viterbi Algorithm. In Proceedings of the 2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Cannes, France, 2–5 July 2019; pp. 1–5. doi:10.1109/SPAWC.2019.8815457.
27. Ye, H.; Li, G.Y.; Juang, B.H. Power of Deep Learning for Channel Estimation and Signal Detection in OFDM Systems. *IEEE Wirel. Commun. Lett.* **2018**, *7*, 114–117. doi:10.1109/LWC.2017.2757490.
28. He, H.; Wen, C.K.; Jin, S.; Li, G.Y. Model-Driven Deep Learning for MIMO Detection. *IEEE Trans. Signal Process.* **2020**, *68*, 1702–1715. doi:10.1109/TSP.2020.2976585.
29. Soltani, M.; Pourahmadi, V.; Mirzaei, A.; Sheikhzadeh, H. Deep Learning-Based Channel Estimation. *IEEE Commun. Lett.* **2019**, *23*, 652–655. doi:10.1109/LCOMM.2019.2898944.

30. Burse, K.; Yadav, R.N.; Shrivastava, S.C. Channel Equalization Using Neural Networks: A Review. *IEEE Trans. Syst. Man Cybern. Part C* **2010**, *40*, 352–357. doi:10.1109/TSMCC.2009.2038279.
31. Mao, H.; Lu, H.; Lu, Y.; Zhu, D. RoemNet: Robust Meta Learning Based Channel Estimation in OFDM Systems. In Proceedings of the ICC 2019—2019 IEEE International Conference on Communications (ICC), Shanghai, China, 20–24 May 2019; pp. 1–6. doi:10.1109/ICC.2019.8761319.
32. Wen, C.K.; Shih, W.T.; Jin, S. Deep Learning for Massive MIMO CSI Feedback. *IEEE Wirel. Commun. Lett.* **2018**, *7*, 748–751. doi:10.1109/LWC.2018.2818160.
33. Yang, Y.; Gao, F.; Li, G.Y.; Jian, M. Deep Learning-Based Downlink Channel Prediction for FDD Massive MIMO System. *IEEE Commun. Lett.* **2019**, *23*, 1994–1998. doi:10.1109/LCOMM.2019.2934851.
34. Arnold, M.; Dorner, S.; Cammerer, S.; Hoydis, J.; ten Brink, S. Towards Practical FDD Massive MIMO: CSI Extrapolation Driven by Deep Learning and Actual Channel Measurements. In Proceedings of the 2019 53rd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 3–6 November 2019; pp. 1972–1976. doi:10.1109/IEEECONF44664.2019.9048863.
35. Jiang, W.; Schotten, H.D. Multi-Antenna Fading Channel Prediction Empowered by Artificial Intelligence. In Proceedings of the 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), Chicago, IL, USA, 27–30 August 2018; pp. 1–6. doi:10.1109/VTCFall.2018.8690550.
36. Gruber, T.; Cammerer, S.; Hoydis, J.; Brink, S.t. On deep learning-based channel decoding. In Proceedings of the 2017 51st Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA, 22–24 March 2017; pp. 1–6. doi:10.1109/CISS.2017.7926071.
37. Cammerer, S.; Gruber, T.; Hoydis, J.; ten Brink, S. Scaling Deep Learning-Based Decoding of Polar Codes via Partitioning. In Proceedings of the GLOBECOM 2017—2017 IEEE Global Communications Conference, Singapore, 4–8 December 2017; pp. 1–6. doi:10.1109/GLOCOM.2017.8254811.
38. He, Y.; Zhang, J.; Jin, S.; Wen, C.K.; Li, G.Y. Model-Driven DNN Decoder for Turbo Codes: Design, Simulation, and Experimental Results. *IEEE Trans. Commun.* **2020**, *68*, 6127–6140. doi:10.1109/TCOMM.2020.3010964.
39. Blanquez-Casado, F.; Aguayo Torres, M.d.C.; Gomez, G. Link Adaptation Mechanisms Based on Logistic Regression Modeling. *IEEE Commun. Lett.* **2019**, *23*, 942–945. doi:10.1109/LCOMM.2019.2903045.
40. Luo, F.L. (Ed.) *Machine Learning for Future Wireless Communications*; John Wiley & Sons Ltd.: Hoboken, NJ, USA, 2020.
41. Liaskos, C.; Nie, S.; Tsioliaridou, A.; Pitsillides, A.; Ioannidis, S.; Akyildiz, I. A New Wireless Communication Paradigm through Software-Controlled Metasurfaces. *IEEE Commun. Mag.* **2018**, *56*, 162–169. doi:10.1109/MCOM.2018.1700659.
42. Elbir, A.M.; Papazafeiropoulos, A.; Kourtessis, P.; Chatzinotas, S. Deep Channel Learning for Large Intelligent Surfaces Aided mm-Wave Massive MIMO Systems. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 1447–1451. doi:10.1109/LWC.2020.2993699.
43. Taha, A.; Alrabeiah, M.; Alkhateeb, A. Enabling Large Intelligent Surfaces With Compressive Sensing and Deep Learning. *IEEE Access* **2021**, *9*, 44304–44321. doi:10.1109/ACCESS.2021.3064073.
44. Thilina, K.M.; Choi, K.W.; Saquib, N.; Hossain, E. Machine Learning Techniques for Cooperative Spectrum Sensing in Cognitive Radio Networks. *IEEE J. Sel. Areas Commun.* **2013**, *31*, 2209–2221. doi:10.1109/JSAC.2013.131120.
45. Lee, W.; Kim, M.; Cho, D.H. Deep Cooperative Sensing: Cooperative Spectrum Sensing Based on Convolutional Neural Networks. *IEEE Trans. Veh. Technol.* **2019**, *68*, 3005–3009. doi:10.1109/TVT.2019.2891291.
46. Wang, Y.; Liu, M.; Yang, J.; Gui, G. Data-Driven Deep Learning for Automatic Modulation Recognition in Cognitive Radios. *IEEE Trans. Veh. Technol.* **2019**, *68*, 4074–4077. doi:10.1109/TVT.2019.2900460.
47. Ahmed, K.I.; Tabassum, H.; Hossain, E. Deep Learning for Radio Resource Allocation in Multi-Cell Networks. *IEEE Network* **2019**, *33*, 188–195. doi:10.1109/MNET.2019.1900029.
48. Ghadimi, E.; Davide Calabrese, F.; Peters, G.; Soldati, P. A reinforcement learning approach to power control and rate adaptation in cellular networks. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017; pp. 1–7. doi:10.1109/ICC.2017.7997440.
49. Sun, H.; Chen, X.; Shi, Q.; Hong, M.; Fu, X.; Sidiropoulos, N.D. Learning to Optimize: Training Deep Neural Networks for Interference Management. *IEEE Trans. Signal Process.* **2018**, *66*, 5438–5453. doi:10.1109/tsp.2018.2866382.
50. Challita, U.; Dong, L.; Saad, W. Proactive Resource Management for LTE in Unlicensed Spectrum: A Deep Learning Perspective. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 4674–4689. doi:10.1109/TWC.2018.2829773.
51. Bonati, L.; d'oro, S.; Polese, M.; Basagni, S.; Melodia, T. Intelligence and Learning in O-RAN for Data-driven NextG Cellular Networks. *arXiv* **2020**, arXiv:2012.01263.
52. Zhou, P.; Fang, X.; Wang, X.; Long, Y.; He, R.; Han, X. Deep Learning-Based Beam Management and Interference Coordination in Dense mmWave Networks. *IEEE Trans. Veh. Technol.* **2019**, *68*, 592–603. doi:10.1109/TVT.2018.2882635.
53. Li, X.; Ni, R.; Chen, J.; Lyu, Y.; Rong, Z.; Du, R. End-to-End Network Slicing in Radio Access Network, Transport Network and Core Network Domains. *IEEE Access* **2020**, *8*, 29525–29537. doi:10.1109/ACCESS.2020.2972105.
54. Abbas, K.; Afaq, M.; Ahmed Khan, T.; Rafiq, A.; Song, W.C. Slicing the Core Network and Radio Access Network Domains through Intent-Based Networking for 5G Networks. *Electronics* **2020**, *9*, 1710. doi:10.3390/electronics9101710.
55. Zappone, A.; Renzo, M.D.; Debbah, M. Wireless Networks Design in the Era of Deep Learning: Model-Based, AI-Based, or Both? *IEEE Trans. Commun.* **2019**, *67*, 7331–7376. doi:10.1109/tcomm.2019.2924010.

56. Hammouti, H.E.; Ghogho, M.; Raza Zaidi, S.A. A Machine Learning Approach to Predicting Coverage in Random Wireless Networks. In Proceedings of the 2018 IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–6. doi:10.1109/GLOCOMW.2018.8644199.
57. Mulvey, D.; Foh, C.H.; Imran, M.A.; Tafazolli, R. Cell Fault Management Using Machine Learning Techniques. *IEEE Access* **2019**, *7*, 124514–124539. doi:10.1109/ACCESS.2019.2938410.
58. Association, G. *An Introduction to Network Slicing*; Technical Report; GSMA Head Office, London, UK, 2017.
59. Available online: https://www.thefastmode.com/self-organizing-network-son-vendors (accessed on 10 October 2021).
60. Kibria, M.G.; Nguyen, K.; Villardi, G.P.; Zhao, O.; Ishizu, K.; Kojima, F. Big Data Analytics, Machine Learning, and Artificial Intelligence in Next-Generation Wireless Networks. *IEEE Access* **2018**, *6*, 32328–32338. doi:10.1109/ACCESS.2018.2837692.
61. *Technical Specification Group Services and System Aspects; Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services*; 3GPP TS 23.288 Version 16.3.0 Release 16; Technical Report; 3GPP Mobile Competence Centre: Sophia Antipolis, France, March 2021.
62. *5G System; Network Data Analytics Services 3GPP TS 29.520 V15.3.0*; Technical Report; 3GPP Mobile Competence Centre: Sophia Antipolis, France, April 2019.
63. *5G System; Network Data Analytics Services 3GPP TS 29.520 V16.7.0*; Technical Report; 3GPP Mobile Competence Centre: Sophia Antipolis, France, March 2021.
64. Potluri, R.; Young, K. Analytics Automation for Service Orchestration. In Proceedings of the 2020 International Symposium on Networks, Computers and Communications (ISNCC), Montreal, QC, Canada, 20–22 October 2020; pp. 1–4. doi:10.1109/ISNCC49221.2020.9297249.
65. Chih-Lin, L.; Sun, Q.; Liu, Z.; Zhang, S.; Han, S. The Big-Data-Driven Intelligent Wireless Network: Architecture, Use Cases, Solutions, and Future Trends. *IEEE Veh. Technol. Mag.* **2017**, *12*, 20–29. doi:10.1109/MVT.2017.2752758.
66. Pateromichelakis, E.; Moggio, F.; Mannweiler, C.; Arnold, P.; Shariat, M.; Einhaus, M.; Wei, Q.; Bulakci, Ö.; De Domenico, A. End-to-End Data Analytics Framework for 5G Architecture. *IEEE Access* **2019**, *7*, 40295–40312. doi:10.1109/ACCESS.2019.2902984.
67. Sevgican, S.; Turan, M.; Gokarslan, K.; Yilmaz, H.B.; Tugcu, T. Intelligent network data analytics function in 5G cellular networks using machine learning. *J. Commun. Netw.* **2020**, *22*, 269–280. doi:10.1109/jcn.2020.000019.
68. *Study on Enablers for Network Automation for the 5G System (5GS)*; Phase 2 (Release 17), 3GPP TR 23.700-91 V17.0.0; Technical Report; 3GPP Mobile Competence Centre: Sophia Antipolis, France, December 2020.
69. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A.Y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial Intelligence and Statistics*; AISTATS: Ft. Lauderdale, FL, USA, 2017.
70. Thapa, C.; Chamikara, M.; Camtepe, S. SplitFed: When Federated Learning Meets Split Learning. *arXiv* **2020**, arXiv:2004.12088.
71. Niknam, S.; Dhillon, H.S.; Reed, J.H. Federated Learning for Wireless Communications: Motivation, Opportunities, and Challenges. *IEEE Commun. Mag.* **2020**, *58*, 46–51. doi:10.1109/MCOM.001.1900461.
72. Jeong, E.; Oh, S.; Kim, H.; Park, J.; Bennis, M.; Kim, S.L. Communication-Efficient On-Device Machine Learning: Federated Distillation and Augmentation under Non-IID Private Data. *arXiv* **2018**, arXiv:1811.11479..
73. Chen, M.; Yang, Z.; Saad, W.; Yin, C.; Poor, H.V.; Cui, S. A Joint Learning and Communications Framework for Federated Learning Over Wireless Networks. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 269–283. doi:10.1109/TWC.2020.3024629.
74. Park, J.; Samarakoon, S.; Elgabli, A.; Kim, J.; Bennis, M.; Kim, S.L.; Debbah, M. Communication-Efficient and Distributed Learning Over Wireless Networks: Principles and Applications. *Proc. IEEE* **2021**, *109*, 796–819. doi:10.1109/JPROC.2021.3055679.
75. Lim, W.Y.B.; Luong, N.C.; Hoang, D.T.; Jiao, Y.; Liang, Y.C.; Yang, Q.; Niyato, D.; Miao, C. Federated Learning in Mobile Edge Networks: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 2031–2063. doi:10.1109/COMST.2020.2986024.
76. Liu, Y.; Yuan, X.; Xiong, Z.; Kang, J.; Wang, X.; Niyato, D. Federated learning for 6G communications: Challenges, methods, and future directions. *China Commun.* **2020**, *17*, 105–118. doi:10.23919/JCC.2020.09.009.
77. ETSI. *Artificial Intelligence and Future Directions for ETSI*; [ETSI White Paper], No. 32; Technical Report; ETSI 06921 Sophia Antipolis CEDEX: Valbonne, France, June 2020.
78. ETSI ENI. Available online: https://www.etsi.org/technologies/experiential-networked-intelligence (accessed on 10 October 2021).
79. *ENI System Architecture, ETSI GS ENI 005 V1.1.1*; Technical Report; ETSI 06921 Sophia Antipolis CEDEX: Valbonne, France, September 2019.
80. *ENI Use Cases, ETSI GS ENI 001 V3.1.1*; Technical Report; ETSI 06921 Sophia Antipolis CEDEX: Valbonne, France, December 2020.
81. ETSI. *ENI Requirements, ETSI GS ENI 002 V3.1.1*; Technical Report; ETSI 06921 Sophia Antipolis CEDEX: Valbonne, France, December 2020.
82. ETSI ENI. Available online: https://eniwiki.etsi.org/index.php?title=Ongoing_PoCs (accessed on 10 October 2021).
83. ETSI. *ENI Vision: Improved Network Experience Using Experiential Networked Intelligence*; [ETSI White Paper No. 44]; Technical Report; ETSI 06921 Sophia Antipolis CEDEX: Valbonne, France, March 2021.
84. ETSI. *Requirements Based on Documented Scenarios ETSI GS ZSM 001 V1.1.1*; Technical Report; ETSI 06921 Sophia Antipolis CEDEX: Valbonne, France, Ocotober 2019.
85. ETSI ZSM. Available online: https://www.etsi.org/technologies/zero-touch-network-service-management (accessed on 10 October 2021).

86. ETSI. *Zero-Touch Network and Service Management [Introductory White Paper]*; Technical Report; ETSI 06921 Sophia Antipolis CEDEX: Valbonne, France, 2017
87. ETSI. *ZSM Reference Architecture ETSI GS ZSM 002 V1.1.1*; Technical Report; ETSI 06921 Sophia Antipolis CEDEX: Valbonne, France, August 2019.
88. ITU-T. Available online: https://www.itu.int/en/ITU-T/focusgroups/ml5g/Pages/default.aspx (accessed on 10 October 2021).
89. [ITU-T Y.3170] *Use Cases for Machine Learning in Future Networks Including IMT-2020, Supplement 55*; Technical Report; Telecommunication Stanrdardization Sector of ITU: Geneva, Switzerland, October 2019.
90. [ITU-T Y.3172] *Architectural Framework for Machine Learning in Future Networks including IMT-2020*; Technical Report; Telecommunication Stanrdardization Sector of ITU: Geneva, Switzerland, June 2019.
91. Wilhelmi, F.; Barrachina-Munoz, S.; Bellalta, B.; Cano, C.; Jonsson, A.; Ram, V. A Flexible Machine-Learning-Aware Architecture for Future WLANs. *IEEE Commun. Mag.* **2020**, *58*, 25–31. doi:10.1109/MCOM.001.1900637.
92. Wilhelmi, F.; Carrascosa, M.; Cano, C.; Jonsson, A.; Ram, V.; Bellalta, B. Usage of Network Simulators in Machine-Learning-Assisted 5G/6G Networks. *IEEE Wirel. Commun.* **2021**, *28*, 160–166. doi:10.1109/MWC.001.2000206.
93. [ITU-T Y.3173]. *Framework for Evaluating Intelligence Levels of Future Networks Including IMT-2020*; Technical Report; Telecommunication Stanrdardization Sector of ITU: Geneva, Switzerland, 2020.
94. ACUMOS. Available online: https://www.acumos.org/ (accessed on 10 October 2021).
95. ITU Challenge. Available online: https://www.itu.int/en/ITU-T/AI/challenge/2020/Pages/PROGRAMME.aspx (accessed on 10 October 2021).
96. ITU-T. Available online: https://www.itu.int/en/ITU-T/focusgroups/an/Pages/default.aspx (accessed on 10 October 2021).
97. *O-RAN: Towards an Open and Smart RAN, [White Paper]*; Technical Report; O-RAN Alliance: Alfter, Germany, October 2018.
98. TIP. Available online: https://telecominfraproject.com/ (accessed on 10 October 2021).
99. Masur, P.; Reed, J. Artificial Intelligence in Open Radio Access Network. *arXiv* **2021**, arXiv:2104.09445.
100. *O-RAN Use Cases and Deployment Scenarios, [White Paper]*; Technical Report; O-RAN Alliance: Alfter, Germany, February 2020.
101. Niknam, S.; Roy, A.; Dhillon, H.S.; Singh, S.; Banerji, R.; Reed, J.H.; Saxena, N.; Yoon, S. Intelligent O-RAN for Beyond 5G and 6G Wireless Networks. *arXiv* **2020**, arXiv:abs/2005.08374.
102. *Open RAN Integration: Run with it, [White Paper]*; Technical Report; O-RAN Alliance: Alfter, Germany, 2021.
103. Warden, P.; Situnayake, D. *Tiny ML. Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2019.
104. Branco, S.; Ferreira, A.G.; Cabral, J. Machine Learning in Resource-Scarce Embedded Systems, FPGAs, and End-Devices: A Survey. *Electronics* **2019**, *8*, 1289. doi:10.3390/electronics8111289.
105. Qualcomm. Available online: https://rebootingcomputing.ieee.org/images/files/pdf/iccv-2019_edwin-park.pdf (accessed on 10 October 2021).
106. Lin, J.; Chen, W.M.; Lin, Y.; Cohn, J.; Gan, C.; Han, S. MCUNet: Tiny Deep Learning on IoT Devices. *arXiv* **2020**, arXiv:abs/2007.10319.
107. David, R.; Duke, J.; Jain, A.; Janapa Reddi, E.A. TensorFlow Lite Micro: Embedded Machine Learning for TinyML Systems. In *Proceedings of Machine Learning and Systems*; Smola, A., Dimakis, A., Stoica, I., Eds.; Systems and Machine Learning Foundation: Indio, CA, USA, 2021; Volume 3, pp. 800–811.
108. Lin, J.; Chen, W.M.; Lin, Y.; Cohn, J.; Gan, C.; Han, S. Benchmarking TinyML Systems: Challenges and Direction. *arXiv* **2020**, arXiv:2007.10319.
109. Tiny ML. Available online: https://www.tinyml.org/ (accessed on 10 October 2021).
110. Association, T.G.I. *European Vision for the 6G Wireless Ecosystem [White Paper]*; Technical Report; The 5G Infrastructure Association: Heidelberg, Germany, 2021.
111. Kaloxylos, A.; Gavras, A.; Camps Mur, D.; Ghoraishi, M.; Hrasnica, H. AI and ML—Enablers for Beyond 5G Networks. *Zenodo* **2020**. doi:10.5281/ZENODO.4299895.
112. Partnership, G.P.P. *5G PPP Phase 3 [Brochure]*; Technical Report; The 5G Infrastructure Public Private Partnership: Heidelberg, Germany, 2021.
113. Windmill. Available online: https://windmill-itn.eu/research/ (accessed on 10 October 2021).
114. *Ariadne Vision and System Concept, [Newsletter]*; Technical Report; The 5G Infrastructure Public Private Partnership: Heidelberg, Germany, January 2021.
115. *Ariadne Deliverable 1.1: ARIADNE Use Case Definition and System Requirements*; Technical Report; The 5G Infrastructure Public Private Partnership: Heidelberg, Germany, January 2021.
116. *Ariadne Deliverable 2.1: Initial Results in D-Band Directional Links Analysis, System Performance Assessment, and Algorithm Design*; Technical Report; The 5G Infrastructure Public Private Partnership: Heidelberg, Germany, January 2021.
117. *5Genesis Deliverable 2.1: Requirements of the Facility*; Technical Report; The 5G Infrastructure Public Private Partnership: Heidelberg, Germany, January 2021.
118. *5Growth Deliverable 1.1: Business Model Design*; Technical Report; The 5G Infrastructure Public Private Partnership: Heidelberg, Germany, January 2021.
119. *Study on Management and Orchestration of Network Slicing for Next Generation Network*; (Release 17), 3GPP TR 28.801 V15.1.0; Technical Report; 3GPP Mobile Competence Centre: France, January 2018.

120. *5Growth Deliverable 2.1: Initial Design of 5G End-to-End Service Platform*; Technical Report; The 5G Infrastructure Public Private Partnership: Heidelberg, Germany, January 2021.
121. *5G-Carmen Deliverable 2.1: 5G-Carmen Use Cases and Requirements*; Technical Report; The 5G Infrastructure Public Private Partnership: Heidelberg, Germany, January 2021.
122. *5G-Carmen Deliverable 5.1: 5G-Carmen Pilon Plan*; Technical Report; The 5G Infrastructure Public Private Partnership: Heidelberg, Germany, January 2021.
123. *Hexa-X Deliverable 1.2: Expanded 6G Vision, Use Cases and Societal Values Including Aspects of Sustainability, Security and Spectrum*; Technical Report; 2021. Available online: https://hexa-x.eu/ (accessed on 10 October 2021).
124. EPSCR. Available online: https://epsrc.ukri.org/research/ourportfolio/researchareas/ait/ (accessed on 10 October 2021).
125. Deep Learning Based Solutions for the Physical Layer of Machine Type Communications. Available online: https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/S028455/1 (accessed on 10 October 2021).
126. Haykin, S. Cognitive radio: Brain-empowered wireless communications. *IEEE J. Sel. Areas Commun.* **2005**, *23*, 201–220. doi:10.1109/JSAC.2004.839380.
127. Secure Wireless Agile Networks. Available online: https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/T005572/1 (accessed on 10 October 2021).
128. Communication Aware Dynamic Edge Computing. Available online: https://www.chistera.eu/projects/connect (accessed on 10 October 2021).
129. Transforming Networks—Building an Intelligent Optical Infrastructure. Available online: https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/R035342/1 (accessed on 10 October 2021).
130. The Mathematics of Deep Learning. Available online: https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/V026259/1 (accessed on 10 October 2021).
131. Graph Neural Networks for Explainable Artificial Intelligence. Available online: https://www.chistera.eu/projects/graphnex (accessed on 10 October 2021).
132. Partnership between Intel and the National Science Foundation. Available online: https://www.nsf.gov/pubs/2019/nsf19591/nsf19591.htm (accessed on 10 October 2021).
133. Partnership between Intel and the National Science Foundation. Available online: https://newsroom.intel.com/wp-content/uploads/sites/11/2020/06/MLWiNS-Fact-Sheet.pdf (accessed on 10 October 2021).
134. NSF New Partnership with Private Industries on Intelligent and Resilient Next-Generation Systems. Available online: https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505904 (accessed on 10 October 2021).
135. NSF New Partnership with Private Industries on Intelligent and Resilient Next-Generation Systems. Available online: https://www.nsf.gov/attachments/302634/public/RINGSv6-2.pdf (accessed on 10 October 2021).
136. Accelercomm. Available online: https://www.accelercomm.com/news/193m-savings-with-improvements-in-5g-radio-signal-processing (accessed on 10 October 2021).